Similar Task Features Shape Judgment and Categorization Processes

Janina A. Hoffmann, Bettina von Helversen, and Jörg Rieskamp

University of Basel, Department of Psychology, Basel, Switzerland

Abstract: 195

Word count: 11,647 (main text)

Author Note. Janina A. Hoffmann, Bettina von Helversen, and Jörg Rieskamp, Department of Psychology, University of Basel, Switzerland. Janina A. Hoffmann is now at the Department of Psychology at the University of Konstanz, Germany. This research was supported by two grants from the Swiss National Science Foundation (100014_130192/1 and 100014_146169/1).

Correspondence concerning this article should be addressed to Janina A. Hoffmann, Department of Psychology, University of Konstanz, Universtätsstraße 10, 78464 Konstanz, Germany. Email: janina.hoffmann@uni-konstanz.de.

Abstract

The distinction between similarity-based and rule-based strategies has instigated a large body of research in categorization and judgment. Within both domains, the task characteristics guiding strategy shifts are increasingly well documented. Across domains, past research has observed shifts from rule-based strategies in judgment to similarity-based strategies in categorization, but limited these comparisons to one prototypical environment, a linear task structure, and a restricted set of strategies. To systematically compare the two domains, we considered several instantiations of rule-based and similarity-based strategies and examined strategy choice across different types of judgment and categorization tasks. Between participants, we varied task characteristics from a one-dimensional linear to a multi-dimensional linear and to two multi-dimensional nonlinear tasks. Irrespective of domain, strategies considered, or model comparison technique used, we find that more participants relied on similarity-based strategies when the functional relationship between the cues and the criterion was nonlinear. Shifts from rule-based strategies in judgment to similarity-based strategies in categorization, however, were rare and most pronounced in one-dimensional environments. These results support the hypothesis that the cognitive strategies people select to solve a judgment or categorization task depend less on the domain but more on the complexity of the task.

Keywords: Judgment; categorization; cognitive processes; strategy selection

On many occasions in everyday life, the same task can demand a coarse classification or a more fine-grained judgment. When hiring a job candidate, for instance, the recruiter may sort the applicants into broad categories such as “qualified” or “unqualified”. Alternatively, the recruiter may judge the applicants’ qualifications on a more fine-grained rating scale from 0 to 100 points; that is, from “not qualified at all” to “highly qualified”. Prototypical tasks used to investigate judgments and categorizations have indeed much in common (Juslin, Olsson, & Olsson, 2003). In addition, both research fields identified two main types of strategies people use to judge or classify objects (Erickson & Kruschke, 1998; Juslin, Olsson et al., 2003; McDaniel, Cahill, Robbins, & Wiener, 2013; von Helversen & Rieskamp, 2008, 2009): similarity-based strategies and rule-based strategies. These strategies make distinct assumptions about the way knowledge is represented and about the cognitive processes underlying judgments and categorizations (Hahn & Chater, 1998; Juslin, Olsson et al., 2003). Whereas similarity-based strategies base inferences on a comparison with concrete instances stored in memory, rule-based strategies rely on explicit abstraction of knowledge (Hahn & Chater, 1998). Given the similarity of the tasks, one might expect that making a coarse or a more fine-grained response does not affect the cognitive process. Yet the two research traditions have mostly described categorizations by similarity-based strategies, whereas judgment processes have been predominantly characterized as rule-based (Juslin, Olsson et al., 2003; von Helversen & Rieskamp, 2009). Confirming this characterization, past research suggests that people frequently shift from rule-based strategies in judgment to similarity-based strategies in categorization (Juslin, Olsson et al., 2003; Pachur & Olsson, 2012; von Helversen, Karlsson, Mata, & Wilke, 2013; von Helversen, Mata, & Olsson, 2010). These strategy shifts are supposed to be driven by task feedback with binary feedback in categorization encouraging similarity-based strategies (Juslin, Olsson et al., 2003).

However, people do not exclusively rely on rule-based strategies in judgment and similarity-based strategies in categorization. Past research has sought to understand which

factors determine strategy choice within each domain by investigating a broad variety of task characteristics (e.g. Ashby & Maddox, 1992; Ashby, Waldron, Lee, & Berkman, 2001; Elwin, Juslin, Olsson, & Enkvist, 2007; Hoffmann et al., 2013; Juslin et al., 2008; McKinley & Nosofsky, 1996; Medin & Schwanenflugel, 1981; Rouder & Ratcliff, 2004; von Helversen & Rieskamp, 2008). Whereas categorization research puts more emphasis on how the number of cues that need to be integrated distinguishes between strategies (Ashby, Maddox, & Bohil, 2002; Erickson & Kruschke, 1998; Maddox & Ashby, 2004; Nosofsky, Palmeri, & McKinely, 1994; Zeithamova & Maddox, 2006), judgment research identified the functional relationship between the cues and the criterion as a major determinant of strategy choice (Hoffmann et al., 2013; Karlsson et al., 2007; Juslin et al., 2008; Olsson, Enkvist, & Juslin, 2006) and gave less consideration to the number of cues. These different foci make it difficult to compare research across domains, which may have concealed common factors underlying strategy choice in both judgment and categorization and thus overestimated differences in strategy choice between the domains. This problem is exacerbated by the fact that the models used to describe similarity- and rule-based processes differ within and between domains. Judgment research typically contrasts linear rules with exemplar models, but the exemplar models considered differ in their complexity (cf. Hoffmann, von Helversen, & Rieskamp, 2014; Pachur & Olsson, 2012; von Helversen et al., 2010). With categorization, in turn, the rules considered are often of a logical nature, constrained to one or two dimensions, and contrasted with exemplar or decision bound models (cf. Ashby et al., 2002; Donkin, Newell, Kalish, Dunn, & Nosofsky, 2015; Filoteo, Lauritzen, & Maddox, 2010; Maddox & Ashby, 2004). Basing strategy classifications on a limited set of strategies can, however, lead to misleading results even within a domain (Donkin et al., 2015). Donkin and colleagues found that the number of people classified to two strategy classes differed greatly as a function of the strategies and the data considered.

In the current work we aim to overcome these problems by systematically manipulating the number of dimensions and the functional relationship across judgment and categorization varying the task domain within participants and the task structure between participants. This systematic and integrative approach allows us to quantify to what extent specific task characteristics encourage strategy choice across domains. Furthermore, it enables us to estimate how far strategy choice is influenced by the task domain across different task structures and how far strategy choice is influenced by preferences for specific strategies across domains. Finally, taking the variety of models proposed in both domains into account we assess the degree to which strategy classification hinges upon specific assumptions about the cognitive processes by comparing the most representative instantiation or a variety of instantiations of rule-based and similarity-based strategies on different model selection criteria. In the following we will review first which rule-based and similarity-based models have been considered in the categorization and judgment literature and the task characteristics that have been identified to influence strategy choice.

Rule-based Strategies in Categorization and Judgment

Rule-based strategies are assumed to base inferences on abstracted knowledge (Juslin et al., 2008; Karlsson, Juslin, & Olsson, 2008). A prototypical rule-based strategy in judgment, the linear model (Juslin, Jones et al. 2003; Juslin, Olsson et al., 2003), assumes that people abstract how each cue relates to the criterion; that is, people try to find out the importance of each cue. The judgment results from the sum of the cue values, weighted by their importance. A recruiter may, for instance, try to figure out which skills, such as programming skills or knowledge of a foreign language, are important for successfully fulfilling the job requirements and assign a high weight to language skills. Accordingly, the recruiter will rate job candidates as more qualified the better they speak the required foreign language. In a similar way, a person may follow the rule that all candidates who prove a certain level of

language skills can be classified as qualified. Hence, the probability of classifying the job candidate as qualified should increase with increasing language skills.

Rule-based strategies proposed in the literature vary in their complexity from rules considering only one or two cues (Ashby & Maddox, 2005; Erickson & Kruschke, 1998; Nosofsky, Little, & Denton, 2011; Nosofsky et al., 1994) to linear rules with several cues (Juslin et al., 2008; Juslin, Olsson et al., 2003; Newell, Weston, Tunney, & Shanks, 2009; Persson & Rieskamp, 2009; Platzer & Bröder, 2013) or even complex nonlinear rules. Are there limits in the complexity rules can take? In judgment the dominant view is that people can and will learn linear additive rules, but have problems with more complex nonlinear rules. Nonlinear rules are assumed to be difficult because they cannot be learnt via a sequential learning process that considers only two subsequent objects (Juslin et al., 2008) and empirical evidence for nonlinear judgment rules is scarce (Brehmer, 1994).

In categorization, it is assumed that people rely on logical rules based on explicit hypothesis testing processes when one or two dimensions need to be considered and are easy to describe verbally (Ashby & Maddox, 2005; Ashby & O'Brien, 2005). When two or more dimensions need to be integrated, it is frequently assumed that people form linear or quadratic decision bounds between two categories via an implicit procedural learning process (Ashby, Alfonso-Reese, Turken, & Waldron, 1998; Ashby & Maddox, 2005; Ashby & O'Brien, 2005), but that the knowledge summarized in decision bounds may not be verbalizable (Ashby & Maddox, 2005; Maddox, Bohil, & Ing, 2004). However, research in linear categorization tasks with multiple cues shows that people build up task knowledge and possess insight into the judgment process (Lagnado, Newell, Kahan, & Shanks, 2006), suggesting that the boundary between explicit and procedural rule-based processes may be less clear cut.[1] Furthermore, the type of rule-based process may change over the learning process: Presumably verbal hypothesis testing processes may dominate in earlier stages, whereas

procedural learning processes take over in later stages (Ashby et al., 2001; Markant & Gureckis, 2014).

Similarity-based Strategies in Categorization and Judgment

Similarity-based strategies, in contrast, are assumed to base inferences on a comparison with concrete instances stored in memory. An exemplar model, a typical similarity-based strategy, assumes that similarity to past instances is used to make a categorization or judgment (Juslin, Olsson et al., 2003; Medin & Schaffer, 1978; Nosofsky, 1988). These exemplar models assume that all previously encountered objects (the exemplars) are stored in long-term memory along with their associated categories (or criterion values in judgment). When categorizing a new object (the probe), past exemplars are retrieved from memory and the probe is compared to all exemplars stored in memory. The more similar the probe is to a past exemplar, the more likely the probe will be classified as belonging to the same category. For instance, when categorizing a new job applicant, recruiters may remind themselves of all employees who have held the job in the past. The more similar the job candidate's language and programming skills are to language and programming skills possessed by qualified past employees, the more likely it is that this candidate is also classified as qualified. Similarly, when judging a new job candidate, the job recruiter may remind himself of all previous employees. The more similar the candidate's skills are to skills of those employees who were judged as highly qualified, the more qualified the candidate will be judged.

Similarity-based strategies vary in the assumptions regarding which information people retrieve and how this retrieval proceeds. Exemplar models varying only in sensitivity assume that people retrieve all exemplars to make a judgment or a categorization with the sensitivity determining whether people retrieve only highly similar exemplars or also more distant ones (Nosofsky & Zaki, 1998). Further, selective attention can modulate which cues people attend to. Exemplar models with attention weighting assume that people learn to

discriminate predictive cues from less predictive ones by focusing attention on the important cues and weighting them more heavily in the final judgment (Nosofsky & Johansen, 2000). More complex exemplar models in categorization finally allow for perceiving cues separately or integrally, for modeling of category biases, as well as for responding more deterministically or by probability matching (Nosofsky & Johansen, 2000; Maddox & Ashby, 1993). Judgment research, in contrast, has mostly focused on exemplar models modeling retrieval specificity (Hoffmann et al., 2013, 2014; von Helversen & Rieskamp, 2008) or attention processes (Mata, von Helversen, Karlsson, & Cüpper, 2012; von Helversen et al., 2010, 2013).

Factors Encouraging Shifts between Rule-based and Similarity-based Strategies

Number of Cues

The categorization literature has suggested that people approach a categorization task by testing simple rules that consider only one or two dimensions. If these rules are not successful, people switch to similarity-based strategies (Erickson & Kruschke, 1998; Nosofsky et al., 1994). For instance, Nosofsky et al. (1994) suggested that people test simple one- or two-dimensional rules when learning categorization tasks, but store exceptions in memory if the rules do not work. Similarly, Erickson and Kruschke (1998) suggested that people simultaneously process rules and exemplars, but restricted the rules tested to one dimension. Furthermore, people seem to process categorization tasks differently if they can be solved by a simple one- or two-dimensional rule compared to categorization tasks that require information integration (Ashby et al., 2002; Filoteo et al., 2010; Maddox & Ashby, 2004; Zeithamova & Maddox, 2006). In sum, this suggests that the number of cues is an important factor driving rule-based or similarity-based strategies in categorization.

In judgment, meta-analyses identified the number of cues as one major factor determining performance (Karelaia & Hogarth, 2008; Kaufmann & Athanasou, 2009). If more cues have to be considered for making a judgment, judgment performance decreases (Karelaia

& Hogarth, 2008). Kareleia and Hogarth (2008) explained this performance decrease by a decreasing match between the linear rules of the judge and the linear model of the environment. If the number of cues increases people may follow more complex strategies instead (Einhorn, 1971) or, alternatively, switch to similarity-based strategies. As a factor influencing strategy choice in judgment, however, the number of cues has — to our knowledge — been neglected.

Functional Relationship between Cues and Criterion

Past research has shown that strategy shifts in judgment are mainly influenced by the functional relationship between the cues and the criterion (Hoffmann et al., 2013; Juslin et al., 2008; Karlsson et al., 2007; von Helversen & Rieskamp, 2008). Indeed, the majority of research suggests that people rely more on similarity-based strategies if the task cannot be solved by a linear rule, for instance, if the criterion is a multiplicative function of the cues (Hoffmann et al., 2013, 2014; Juslin et al., 2008). Quadratic task structures (environments) in which the criterion is a non-linear quadratic function of the cues may represent an exception because the same criterion value is associated with multiple, but dissimilar exemplars. Hence, neither similarity-based strategies nor rule-based linear models yield good performance early in training and people may drop back to the default, but useless rule-based strategy (Karlsson et al., 2008; Olsson et al., 2006).

The functional relationship between cue and criterion has also been studied in function learning tasks in which people learn to predict a continuous criterion based on one continuous cue with varying functional relationships between cue and criterion. Overall, this research suggests that linear functions are learnt faster than exponential or quadratic functions (Busemeyer, Byun, DeLosh, & McDaniel, 1997; DeLosh, Busemeyer, & McDaniel, 1997). In addition, rule-based function learning models fare well on extrapolation for linear functions (that is, at predicting response values outside the training range) but fail on extrapolation for exponential or quadratic functions (De Losh et al., 1997; McDaniel & Busemeyer, 2005).

Associative, similarity-based models, in contrast, account successfully for extrapolation for exponential or quadratic functions if they incorporate a linear extrapolation mechanism (De Losh et al., 1997; McDaniel & Busemeyer, 2005), suggesting a combination of both processes.

In categorization, nonlinear or quadratic category bounds are learned more slowly and less accurately than linear category bounds (Ashby & Gott, 1988), but people can reach near-optimal performance when learning nonlinear bounds (Ashby & Maddox, 1992). How the functional relationship affects rule-based and similarity-based categorization strategies is, however, less clear: Linear environments are accurately described both by deterministic exemplar models and decision bound models (Maddox & Ashby, 1993), but decision bound models fit participants' categorizations worse if the stimuli are integral; that is, perceived holistically, or the decision boundary is oblique and considers two dimensions (McKinley & Nosofsky, 1996). Further, in nonlinear environments decision bound models provide a better description of the data (Maddox & Ashby, 1993), but if the optimal decision bound is more complex than a quadratic bound, exemplar models seem to describe categorizations better (McKinley & Nosofsky, 1995).

In sum, judgment research provides some evidence for strategy shifts towards similarity-based processes with nonlinear functional relationships, but the evidence in function learning and categorization is less clear. In addition, the models tested against each other differ between domains as well as from models considered in studies that investigate the effect of the number of cues, leaving open the question of how much the results depend on specific assumptions underlying these models.

Judgment versus Categorization: Feedback and Individual Preferences

The task descriptions and the strategies in categorization and judgment resemble each other closely (Juslin, Olsson et al., 2003). In both tasks, the objects are described by several cues that predict either a binary categorization or a continuous judgment. People learn to make these continuous judgments (or binary categorizations) over time by receiving

continuous (or binary) feedback about the correct outcome. Likewise, both domains assume that responses in these tasks are based on rule abstraction or similarity-based retrieval (Hoffmann et al., 2014; Juslin et al., 2008; Juslin, Jones et al., 2003; Karlsson et al., 2007; Rouder & Ratcliff, 2004). These similarities suggest that people rely on the same process within both domains. In this vein, it has been argued that people have stable individual preferences for relying on rules or similarity and exhibit those preferences across different domains, unless the task strongly favors one solution (McDaniel et al., 2013). Specifically, McDaniel et al. (2013) found that people identified as rule-learners in linear V-shaped function learning tasks transferred their preferences to a subsequent abstract categorization task.

Past research has also found some striking differences between judgment and categorization. Several studies report that in linear tasks more people rely on a rule-based strategy in judgment, whereas in categorization the majority relies on similarity-based strategies (Juslin, Olsson et al., 2003; Karlsson et al., 2008; von Helversen et al., 2010). One reason why different strategies may emerge across domains is the differential nature of feedback in judgment and categorization (Juslin, Olsson et al., 2003). Task feedback allows the learner to adapt strategies to the task demands over time (Kämmer, Gaissmaier, & Czienskowski, 2013; Rieskamp & Otto, 2006; but see Bröder et al., 2013). Feedback in judgment, however, is more fine-grained than the binary feedback in categorization so that it should facilitate inferring the cue-criterion relationship and ultimately promote the abstraction of rules independent of any other task characteristics (Juslin, Olsson et al., 2003). In this vein, it has been found that if less effort needs to be invested in abstracting cue weights, for instance, when cue directions are known, more people rely on rule-based strategies (Newell et al., 2009; Platzer & Bröder, 2013; von Helversen et al., 2013). Likewise, learning which of two objects has a higher criterion value enhances reliance on rule-based strategies, possibly

because people focus on how differences in cue values are associated with differences in judgment criteria — an important step for rule abstraction (Pachur & Olsson, 2012).

Past research has mostly restricted the comparison of judgment and categorization strategies to one prototypical environment, namely a linear task structure with multiple binary cues, and focused on a restricted set of strategies (but see Pachur & Olsson, 2012). However, the argument that people try to abstract explicit representations of the cue-criterion relationship when provided with informative feedback does not presuppose a linear relationship (Juslin, Olsson et al., 2003), suggesting that more informative feedback may also enable people to develop explicit representations of cue-criterion relationships in more complex tasks. In sum, the degree to which people prefer the same strategy across domains or switch between strategies depending on the domain is still unclear.

The Present Research: Strategy Choice in Judgment and Categorization Across Environments

The current research examines how the domain affects strategy selection across a range of rule-based and similarity-based strategies and across different task structures. Our goal is threefold: First, we aim to assess the degree to which the number of cues and the functional relationship between cue and criterion — task characteristics that have been shown to strongly influence strategy selection in one domain but have been neglected in the other domain — affect strategy selection similarly in both domains. Second, we aim to investigate to what extent people rely more heavily on rule-based strategies in judgment than in categorization and to what extent a stronger focus on rules is influenced by the underlying task structure and personal preferences for strategies. Third, we aim to assess to what extent different methodological approaches, such as different model selection methods or the test of strategy classes versus single strategies, affect the research results.

To this goal, we conducted two experiments. In both experiments participants solved both a categorization and a multiple-cue judgment task with the same underlying task

structure (i.e. environment). Both tasks consisted of a training phase, in which participants first learned to judge (or categorize) a range of objects, and a test phase, in which participants judged (or categorized) new objects. In Experiment 1 we varied the environment on three levels between participants, comparing a one-dimensional linear task (OLIN), a multi-dimensional linear task (MLIN) and a multi-dimensional multiplicative task (MMULT), and task domain (judgment vs. categorization) within participants. Experiment 2 extended these results to a multi-dimensional quadratic task (MQUAD). Because the experimental method varied only slightly between Experiment 1 and 2, we present the results together for the sake of clarity. In a first step we test participants' strategy choices using the most prominent representative from the class of similarity-based and rule-based strategies. In a second step, we investigate how stable those results are when considering a variety of instantiations of rule-based and similarity-based strategies and different model selection criteria.

Method

Participants. In Experiment 1, 96 participants (76 females, M_{Age} = 23.7, SD_{Age} = 5.9) were recruited from the University of Basel. Participants received course credit or a book voucher (worth 25 Swiss Francs, CHF). In addition, they could earn a bonus of 3 CHF in each task and had the opportunity to win one of six vouchers for an Internet-based retailer (worth 25 CHF each). In Experiment 2, 32 participants (25 females, M_{Age} = 26.5, SD_{Age} = 10.7) were recruited from the same participant pool. Participants from Experiment 1 were not allowed to take part in Experiment 2. Participants received course credit or a participation fee (20 CHF per hour). As in Experiment 1, they could earn a bonus of 3 CHF in each task and had the opportunity to win one of two vouchers for an Internet-based retailer (worth 25 CHF each).

Design and material. We used two different cover stories for the categorization and the multiple-cue judgment task. One cover story asked participants to judge the toxicity of a bug: In the multiple-cue judgment task, participants estimated how toxic a bug was on a scale

from 0 to 50, whereas in the categorization task participants classified the bug as toxic or harmless. The other cover story asked participants to judge how successful comic figures, the Sonics, were at catching small animals: In the multiple-cue judgment task, participants judged how many small animals the Sonic caught on a scale from 0 to 50, whereas in the categorization task they classified the Sonic as catching few or many animals.

The stimuli for the two cover stories consisted of pictures of either bugs or Sonics. These bugs and Sonics varied on four different quantitative cues. The bugs varied on the length of their legs, their antennae and their wings, and the number of spots on their back. The Sonics had different sizes of their ears and their nose, and a different number of hairs and stripes on their shirt. These pictorial cues could be used to predict the criterion (the toxicity of a bug or the success of the Sonic).

To manipulate the number of cues and the functional relationship, we varied how these cues had to be combined to form the judgment criterion. In the MLIN environment, the criterion c_{MLIN} was a linear, additive function of the cues:

$$c_{\text{MLIN}} = 4\,x_1 + 3\,x_2 + 2\,x_3 + x_4, \tag{1}$$

where c_1 to c_4 are the cue values ranging from 0 to 5. According to the cue weights, c_1 reflects the most important cue and c_4 the least important one.

In the OLIN environment only one cue predicted the judgment criterion y_{OLIN}:

$$c_{\text{OLIN}} = 10\,x_3. \tag{2}$$

In the MMULT environment, the function generating the criterion y_{MMULT} included a multiplicative combination of the cues:

$$c_{\text{MMULT}} = \frac{4x_1+3x_2+2x_3+x_4+2x_1x_2x_3+x_2x_3x_4}{8.5}. \tag{3}$$

Finally, in the MQUAD environment, the judgment criterion was a quadratic function of the cues:

$$c_{\text{MQUAD}} = 0.83[4(x_1 - 2.5)^2 + 3(x_2 - 2.5)^2 + 2(x_3 - 2.5)^2 + (x_4 - 2.5)^2 - 2.5] \tag{4}$$

Subtracting 2.5 from each cue centered the cue values on their mean. Consequently, high and low cue values are associated with higher criterion values, whereas intermediate cue values are associated with lower criterion values.

In the categorization tasks, the criterion was no longer continuous, but binary. This binary criterion was created by a median split on the corresponding judgment criterion for all possible items. Sonics (or bugs) with criterion values above the median were classified as catching many animals (or as toxic). Sonics (or bugs) with criterion values below the median were classified as catching few animals (or as harmless). This median split creates a linear category boundary in the OLIN and the MLIN environments, and a nonlinear category boundary in the MMULT environment. In the MQUAD environment, the category boundary was spherical; accordingly, the less similar an exemplar is to the prototypical exemplar with intermediate cue values, the more likely it is that the exemplar belongs to a different category than the prototype.

For each participant, the cues x_1 to x_4 were randomly assigned to the pictorial cues (e.g., ears or nose). Higher cue values, however, were always associated with more salient pictorial features. For instance, a cue value of zero corresponded to a bug without spots on the back and a cue value of five to a bug with five spots on its back. Likewise, a cue value of zero on the cue "legs" corresponded to a bug without (visible) legs, whereas a bug with a cue value of five had long legs.

From all possible items, we constructed two different item sets: a training set and a validation set. The training set was used in the training phase to allow participants to learn how to solve the tasks. The validation set was employed in the test phase to identify the judgment or categorization strategy that people followed. For constructing the item sets we first generated 1000 trainings sets, each consisting of 25 training items. Second, we selected one training set fulfilling two criteria: (a) For categorizations, one- or two-dimensional rules should not lead to a high accuracy in the multidimensional environments. (b) Linear rule-

based strategies should solve the judgment and categorization task in the MMULT environment worse than similarity-based strategies; that is, a linear model fitted the training set worse than an exemplar model with a sensitivity parameter. Next, we generated 100 validation sets consisting of 15 validation items and selected one validation set that strongly discriminated between the predictions of a linear rule-based model and an exemplar model with one sensitivity parameter for the judgment and categorization task in each environment. Table 1 depicts the final training set and Table 2 the validation set for Experiment 1. We used the same selection criteria as for the MMULT environment to create the training-validation set combination for Experiment 2 (see Table 3 and Table 4 for the specific items used).

Procedure. In Experiment 1, we assigned 32 participants randomly to each of the three environments (OLIN, MLIN, or MMULT). In Experiment 2, all 32 participants solved a MQUAD environment. The assignment of the cover stories to the judgment and categorization task and the order of the two tasks were counterbalanced within each environment.

During the training phase, participants learned to predict the criterion value (or the category) of 25 training items. In each trial, they first estimated the criterion or categorized the item. Afterwards they received feedback about their own answer, the correct outcome, and the points they earned. In a training block, all 25 training items were presented in random order. The training phase ended after 10 training blocks and participants moved on to the test phase. In this test phase, participants judged all 15 new validation items four times without receiving any feedback.

Participants were incentivized to achieve a high task performance. In each trial of the categorization task, participants could earn 20 points for a correct answer, 10 points for items that were classified with a probability of .5 to both categories, and 0 points for an incorrect answer. In the judgment task, participants earned more points the less their judgment j deviated from the correct criterion y:

$$\text{Points} = 20 - \frac{(j - y)^2}{7.625} \quad (4)$$

This function was truncated so that participants could win at most 20 points and could not lose any points in each trial. The more points participants earned in a task, the higher their chances of winning a retailer voucher for that task. In addition, participants could earn a bonus of 3 CHF in each task if they reached 80% of the points in the last training block. In the categorization task, this learning criterion corresponded to 80% correct classifications. In the judgment task, judgment accuracy was measured in root-mean-square deviations (RMSD) between participants' judgments and the criterion. Participants reached the learning criterion if judgment accuracy was below 5.5 RMSD in the last training block. In Experiment 2, we relaxed the learning criterion for the judgment task to strongly encourage participants to learn the task: Participants could earn a bonus of 3 CHF if they reached more than 55% of the points in the last training block in the judgment task. The relaxed learning criterion in the judgment task corresponds approximately to a RMSD below 10 and accordingly participants reaching the learning criterion should still outperform a linear rule-based model and a guessing model by 2 RMSD ($RMSD_{Guessing} = 12$, $RMSD_{Linear} = 11.8$).

Results: Single Strategy, Single Method

In the following we first report participants' accuracies in categorization and judgment, separately for the two domains. Next, we describe the cognitive models used to describe participants' judgment and categorization strategies and how task domain and environment impact on strategy selection. Finally, we analyze the influence of individual preferences on strategy selection across the different environments. Because accuracy in judgment and categorization as well as measures of model fits were not normally distributed we relied on non-parametric statistical tests.

Accuracy in the judgment task. Table 5 reports judgment accuracy in the last block of the training phase and in the test phase. Judgment accuracy in training was measured as the RMSD between the criterion value and participants' judgment in the last training block; to determine judgment accuracy in the test phase, we averaged participants' judgments over the four presentations and calculated the RMSD between the criterion and the averaged judgments. At the end of training participants made more accurate judgments in the OLIN environment than in the MLIN environment ($U = 200$, $p < .001$) or in the MMULT environment ($U = 222$, $p < .001$), but judgment accuracy did not differ between the MLIN and the MMULT environments ($U = 633$, $p = .106$). The cover story (bugs or Sonics) did not affect judgment accuracy ($U = 1132$, $p = .886$), but participants solved the judgment task slightly better when the categorization task was solved first (RMSD = 4.6) than second (RMSD = 6.3, $U = 1405.5$, $p = .063$). Also at test, judgment accuracy was higher in the OLIN environment than in the MLIN environment ($U = 131$, $p < .001$) or in the MMULT environment ($U = 159$, $p < .001$), but did not differ between the MLIN and the MMULT environments ($U = 571$, $p = .435$).

In the MQUAD environment in Experiment 2, participants made on average less accurate judgments than in the MLIN and MMULT environments from Experiment 1 in both training and test. However, 19 participants (59.4%) still outperformed a guessing model in the last training block. The order of the tasks did not affect judgment accuracy in the last training block, $F(1,29) = 0.79$, $p = .382$, but participants were slightly better at judging bugs (RMSD = 10.8, $SD = 2.4$) than Sonics in the last training block (RMSD = 13.1, $SD = 3.0$), $F(1,29) = 5.74$, $p = .023$.

Accuracy in the categorization task. Overall, participants performed better in the OLIN environment than in the MLIN or the MMULT environments. Table 5 reports the mean percentage of errors in the last training block and the four test blocks. Participants made fewer errors in the OLIN environment than in the MLIN or MMULT environment in the last

training block as well as in the test phase. Like in the judgment task, participants made fewer errors at the end of training in the OLIN environment than in the MLIN environment (Mann-Whitney $U = 64.5$, $p < .001$) and in the MMULT environment ($U = 85.5$, $p < .001$), but there was no difference in the error rates between the MLIN and the MMULT environments ($U = 499.5$, $p = .871$). The order of the tasks (categorization or judgment task first) and the cover story (bugs or Sonics) did not affect how well people learned the task ($U = 1101$, $p = .709$, and $U = 1032.5$, $p = .379$, respectively). Similarly, participants made fewer errors during the test phase in the OLIN environment than in the MLIN environment ($U = 47$, $p < .001$) and in the MMULT environment ($U = 60.5$, $p < .001$). However, error rates did not differ between the MLIN and MMULT environments ($U = 441.5$, $p = .347$). Taken together, these performance results suggest that the number of cues affected how well people learned to solve the categorization and the judgment task but not the functional relationship.

In Experiment 2, participants categorized the objects less accurately in the MQUAD environment than in the MLIN or MMULT environment from Experiment 1. Despite the lower accuracy in training, 25 participants (78.1%) still outperformed a guessing model predicting 44% of errors in the training phase. Categorization accuracy in the last training block was affected neither by order, $F(1,29) = 2.15$ $p = .153$, nor by the cover story, $F(1,29) = 0.14$, $p = .716$. Similarly, participants made more errors in the test phase than in Experiment 1.

Modeling of cognitive processes. To identify the cognitive strategies that people rely on in judgment and categorization, we first used a computational modeling approach that has often been employed to study judgment strategies (Hoffmann et al., 2013, 2014; von Helversen & Rieskamp, 2008). This single-strategy approach contrasts one prominent representative of the class of rule-based strategies, the linear model (Lin), with one prominent representative of the class of similarity-based strategies, an *Ex*emplar model with one sensitivity parameter but *no* parameters for the *Att*ention given to each dimension (ExNoAtt, see Appendix A for more details on model specification, estimation, and selection). The linear

model, corresponding mathematically to a linear regression model, has often served as the prototypical rule-based strategy in judgment tasks (Cooksey, 1996). It can represent simple rule-based strategies relying on a single cue, but also allows more complex rules combining several cues in a linear additive fashion. However, it does not include nonlinear rules or interactions. The ExNoAtt model, a prototypical similarity-based model, compares the present object to all previously encountered exemplars. The more the cue values match each other, the smaller is the distance between the objects and, hence, the objects are more similar to each other (Nosofsky & Johansen, 2000). Objects with higher similarity influence the final judgment (or category bias) more strongly. The ExNoAtt model allows the modeling of how specifically people retrieve the objects but it does not assume that people focus attention on specific cues. Finally, we included a random guessing model that assumes that participants' responses vary randomly around a mean on each trial (Maddox et al., 2009).

To select the model that best describes each participant, we used a generalization test (Busemeyer & Wang, 2000). We fitted each model to participants' judgments (or categorizations) from the last three training blocks and predicted participants' responses on validation items in the test phase. This generalization test corrects not only for model complexity in terms of the number of free parameters, but also for functional complexity (Busemeyer & Wang, 2000). We then determined relative model fit by calculating the deviances for each model based upon the difference between model predictions and participants' responses. We weighted these deviances according to their success in predicting the responses. Deviance weights close to 1 for a model indicate that the model is more likely to have generated the data than the other models in the set, whereas deviance weights close to 0 for a model indicate that the model is less likely to have generated the data.

Model fits in judgment. The left upper panel in Figure 1 shows average deviance weights for the guessing, rule-based, and similarity-based model in the judgment task, separately for the four different environments (see Table 6 for descriptive statistics).

Descriptively, deviance weights for the rule-based model decline from the OLIN to the MLIN environment and are again slightly lower in the MMULT environment, whereas average weights for the similarity-based model increase. In the MQUAD environment, deviance weights are similar for the rule-based and similarity-based model, but slightly increase for the guessing model. The rule-based model outperformed the guessing model in all environments (all $p < .001$, $r < -.63$), but the similarity-based model could not be distinguished from guessing in the OLIN and the MQUAD environment (OLIN: $p = .086$, $r = -.30$; MLIN: $p < .001$, $r = -.91$; MMULT: $p < .001$, $r = -.72$; MQUAD: $p = .060$, $r = -.33$). Further, the rule-based model predicted participants' judgments better than the similarity-based model only in the OLIN environment ($p < .001$, $r = -.94$), but not in the MLIN ($p = .052$, $r = -.34$), MMULT ($p = .963$, $r = -.01$), or the MQUAD environment ($p = .340$, $r = -.17$). In sum, the relative model fit for the validation items in the test phase (i.e. the deviance weights) successfully identified the best model for the OLIN task; however, for all other tasks the relative model fit did not allow to uniquely identify the best model.

Model fits in categorization. The left lower panel in Figure 1 shows average deviance weights for the guessing, rule-based, and similarity-based model in the categorization task, separately for the four different environments (see Table 7 for descriptive statistics). Descriptively, deviance weights for the rule-based model decline from the OLIN to the MLIN environment, but do not change in the MMULT environment. Similarly, average weights for the similarity-based model increase from the OLIN to the MLIN environment, but not more in the MMULT environment. In the MQUAD environment in Experiment 2, deviance weights are finally higher for the similarity-based model than for the rule-based or the guessing model. The rule-based model outperformed the guessing model in all environments in Experiment 1 (all $p < .001$, $r < -.68$), but not in the MQUAD environment in Experiment 2 ($p = 1.00$, $r = .00$). The similarity-based model made more accurate predictions than the guessing model in all environments (all $p < .001$, $r < -.57$). In the OLIN environment,

the rule-based model still outperformed the similarity-based model ($p < .001$, $r = -.72$), but model fits could not distinguish the models in the MLIN ($p = .304$, $r = -.18$) and the MMULT environment ($p = .934$, $r = -.01$). In the MQUAD environment from Experiment 2, the similarity-based model fared better than the rule-based model in predicting participants' categorizations ($p = .008$, $r = -.47$).

Predicting strategy choice. To further investigate how the task environment and the type of task changed strategy choice, we classified participants to the rule-based, the similarity-based and the guessing model. Overall, only a minority of participants was assigned to the guessing model in the categorization or the judgment task in Experiment 1 (see Table 6 for strategy classifications in judgment and Table 7 for categorization), whereas a guessing model best described more participants in the MQUAD environment from Experiment 2. Descriptively, strategy classifications show a pattern similar to the average deviance weights within the task domains. Comparing classifications between domains across all environments shows that a few more participants were classified to the rule-based model in judgment ($n = 83$) than in categorization ($n = 71$). Conversely, slightly more participants were assigned to the similarity-based model in categorization ($n = 51$) than in judgment ($n = 40$).

Figure 2 (judgment task) and Figure 3 (categorization task) illustrate how well the predictions of the two models match participants' responses as well as differences in the model predictions. The upper rows show predictions of the rule-based linear model (white crosses) for participants classified to the linear model (black diamonds) and the ExNoAtt model (gray circles) and the lower rows show predictions of the similarity-based ExNoAtt model. Overall, model predictions match the behavior of the participants classified to the respective model quite well and indicate where the models make different predictions. Particularly, the ExNoAtt model predicts a larger variability of responses in the OLIN and the MLIN environment than the linear model, whereas the linear model predicts in the MMULT environment an overestimation of low criterion values in judgment and an underestimation of

moderate criterion values in categorization. In the MQUAD environment the linear model predicts in both judgment and categorization that participants' responses are independent from the optimal criterion value, whereas the ExNoAtt model predicts that participants can learn to a small extent to adapt their responses to the optimal criterion.

To investigate how the number of cues and the functional relationship influenced judgment and categorization strategies, we collapsed the data from both experiments and conducted a mixed logistic regression analysis on the assigned strategy from the generalization test, excluding participants classified to the guessing model separately for each task. We included the task domain and the environment as fixed factors, setting for the environment one contrast comparing the linear environments against the nonlinear environments (OLIN and MLIN against MMULT and MQUAD), one comparing the number of cues in the linear environment (OLIN against MLIN), and one comparing the two nonlinear environments (MMULT against MQUAD). In addition, we included for participants a random intercept. To select the most important predictors for strategy selection, we compared models with different predictors using a likelihood ratio test and also report Akaike's Information criterion (AIC).

In comparison to a random model that only includes the random intercept for participants (AIC = 323) as a predictor, a model additionally using the environment as a predictor (AIC = 291) fared better, χ^2 (3) = 37.7, $p < .001$, whereas using the task domain (AIC = 322) as a predictor did not significantly improve model fit, χ^2 (1) = 2.8, $p = .095$. Likewise, accounting for the interaction between environment and task domain did not improve model fit, AIC = 289, χ^2 (3) = 7.1, $p = .069$. The contrasts on the different environments suggest that more people are classified to similarity-based strategies in nonlinear than in linear environments, OR = 2.3, CI = [1.6; 3.4] with confidence intervals based on the likelihood-ratio method. In linear environments, more cues also increase the number of participants classified as relying on similarity-based strategies OR = 2.8, CI = [1.7;

5.3], but classifications in the nonlinear environments MMULT and MQUAD do not differ from each other, OR = 1.3, CI = [0.8; 2.0]. Overall, these results suggest that categorization tasks do not increase reliance on similarity-based strategies, whereas the number of cues and a nonlinear functional relationship increase reliance on similarity-based strategies.

Strategy preferences and adaptation. Investigating rule-based and similarity-based strategies across two domains allows us to investigate to what extent people's strategy choices depend on an individual preference for one type of strategy (McDaniel et al., 2013) or result from an adaptation to the task demands (Rieskamp & Otto, 2006). Specifically, the individual preferences hypothesis suggests that if people relied on a similarity-based strategy in the first task, then the conditional probability of using a similarity-based strategy in the subsequent task should also be high. Likewise, if a person preferred a rule-based strategy in the first task, the conditional probability of using a rule-based strategy in the second task should also be high. The adaptation hypothesis, in contrast, predicts that the magnitude of this conditional probability depends on the environment: In tasks that can be solved by linear rules only the conditional probability for rules should be high, whereas in tasks that can better be solved by the exemplar model, only the conditional probability for an exemplar model should be high.

To find out whether people possessed stable individual preferences across tasks or whether they adapted their strategy in response to feedback, we classified participants in a first step as following a linear model or an exemplar model in both tasks or as shifting between strategies in both tasks irrespective of the type of task (judgment or categorization). Descriptively, in the OLIN environment most participants (n = 26) relied upon a linear model in both tasks, but the number of participants following a linear model decreased to the MLIN (n = 14) to the MMULT (n = 9) to the MQUAD environment (n = 4). By contrast, the number of participants assigned to the exemplar model in both tasks increased from the OLIN (n = 0) to the MLIN (n = 6) to the MMULT (n = 8) to the MQUAD environment (n = 9). Likewise, the number of participants shifting between the guessing, linear, and exemplar model also

increased across environments (OLIN: $n = 6$; MLIN: $n = 12$; MMULT; $n = 15$; MQUAD: $n = 19$).

The left graph in Figure 4 depicts the conditional probability of following a rule-based linear model (a similarity-based ExNoAtt model) in the second task given that participants were best described by a linear model (or a ExNoAtt model, respectively) in the first task. If half of the participants shifted from the linear model in the first task to a ExNoAtt or a guessing model in the second task, a probability of .5 would be expected. In the OLIN environment, participants were likely to stay with the linear model in the second task if they were best described by it in the first task, $p(\text{Rule}_{\text{Second}}|\ \text{Rule}_{\text{First}})$. In addition, they were unlikely to follow the ExNoAtt model in the second task, even if they were best described by it in the first task, $p(\text{Similarity}_{\text{Second}}|\ \text{Similarity}_{\text{First}})$. $p(\text{Rule}_{\text{Second}}|\ \text{Rule}_{\text{First}})$ decreased from the MLIN environment to the MMULT environment and dropped below .5 in the MQUAD environment, suggesting that participants were unlikely to follow a rule-based strategy in the second task even if they used it in the first task. In contrast, $p(\text{Similarity}_{\text{Second}}|\ \text{Similarity}_{\text{First}})$ increased from the OLIN to the MLIN environment, but did not vary between the MLIN and MMULT environment. In the MQUAD environment, $p(\text{Similarity}_{\text{Second}}|\ \text{Similarity}_{\text{First}})$ was higher than .5, whereas $p(\text{Rule}_{\text{Second}}|\ \text{Rule}_{\text{First}})$ was close to 0 indicating that participants tended to stay with the ExNoAtt model in the second task, but not with the linear model. Taken together, the conditional probabilities $p(\text{Rule}_{\text{Second}}|\ \text{Rule}_{\text{First}})$ and $p(\text{Similarity}_{\text{Second}}|\ \text{Similarity}_{\text{First}})$ change across environments and, more importantly, strongly differ from each other. These results disagree with the idea that people possess stable preferences for rule-based or similarity-based strategies and provide more evidence for the idea that people adapt the cognitive strategy to the task demands.

It is possible that people change strategies to better solve the task and hence a strategy switch should lead to more accurate judgments and categorizations. To test this idea, we investigated whether strategy shifts can be linked back to task performance. Specifically, we

tested whether participants who shifted strategies in the second task show a stronger learning effect than those who did not switch strategies. Because these strategy shifts are rare, we analyzed all environments together. To only consider relative performance improvements, we z-standardized judgment and categorization performance (measured in RMSD or categorization error, respectively) separately for each environment and task domain. We then tested whether strategy shifts improved performance comparing this model to a random model and a model considering that performance may improve from the first to the second task. Compared to a model with participant as a random effect (AIC = 724), a model including the effect of task order (AIC = 717) fares better, $\chi^2(1) = 8.7$, $p = .003$. Including strategy shifts further improves model fit, AIC = 715, $\chi^2(1) = 4.1$, $p = .04$, but not adding the interaction between task order and strategy shifts, AIC = 715, $\chi^2(1) = 1.8$, $p = .181$. Overall, this analysis suggests that all participants show a learning effect from the first to the second task, $b = -0.35$, $t(127) = -3.0$, $p = .004$, but those who switched strategies between tasks had on average a lower performance than those who did not switch, $b = 0.26$, $t(126) = 2.0$, $p = .046$.

Summary and Discussion: Single Strategy, Single Method

Experiment 1 and 2 studied strategy selection varying the task structure as well as the task domain. Across judgment and categorization, we found that more participants relied on linear rules in linear than in nonlinear tasks, whereas more participants were better described by an exemplar model in nonlinear tasks. When more dimensions had to be integrated linearly, the number of participants best described by rule abstraction decreased. Further, our results on strategy choice do not strongly support the idea that the effectiveness of feedback depends on the environment. Although, across all environments, a few participants were better described by a similarity-based exemplar model in categorization than in judgment, we did not find a major strategy shift between judgment and categorization. Lastly, investigating how consistently people chose rule-based or similarity-based strategies across the two domains

suggested that the conditional probability for choosing one strategy in the second task, given that participants chose the strategy in the first task, varied strongly between rule-based and similarity-based strategies and changed depending on the environment. Switching strategies between tasks, however, did not improve performance relative to those participants who did not switch. In combination, these results suggest that people do not stick with their preferred strategy across domains if it is maladaptive, but switching to another strategy does not help to catch up with those participants who continued using the same strategy.

In the analysis reported above, we focused only on two strategies from each strategy class, rule-based and similarity-based, that have been repeatedly contrasted in judgment research (Hoffmann et al., 2013, 2014; von Helversen & Rieskamp, 2008): a linear, additive model and an exemplar model with one sensitivity parameter. These strategies we considered make rather strict assumptions about the rules people form or how people distribute attention. In our analyses, for instance, the exemplar model was restricted to pay equal attention to all dimensions. Exemplar models applied to categorization data, however, typically use varying attention weights to reflect the idea that important dimensions attract attention, whereas less important dimensions are less attended to (Nosofsky & Johansen, 2000). Similarly, rule-based strategies may take more complex forms than a linear, additive rule. For instance, decision bound models in categorization assume that people can also form quadratic bounds between two categories (Maddox & Ashby, 1993). In this vein, Donkin et al. (2015) presented results suggesting that considering only a subset of models may change the number of participants classified to one strategy class. Taken together, it is possible that the lack of significant differences between judgment and categorization resulted from neglecting more complex rule-based and similarity-based strategies.

In addition, strategy classifications may depend not only on the strategies considered, but also on the data and the model comparison techniques used (Donkin et al., 2015). In the analysis reported above, we relied on the generalization test to classify participants. The

generalization test not only punishes overly complex models for the number of parameters, but also accounts for functional complexity (Busemeyer & Wang, 2000). However, estimating model parameters based on a reduced set of data may leave out parameter values that account for the full range of data, particularly when fitting more complex exemplar models. Accordingly, in the next section, we present to what extent our results vary when considering a larger variety of rule-based and similarity-based strategies as well as different model comparison methods.

Comparison of strategy instantiations and model selection techniques

To estimate the impact that variations in the strategies considered and the model comparison technique have on strategy classification, we considered a range of instantiations of the rule-based and similarity-based strategy class with varying complexity. In the rule-based strategy class, we allowed for more complex rules than linear ones by including nonlinear interactions between cues (*Mult* using two interaction terms and the reduced versions *Mult1* and *Mult2* using one interaction term each) or a quadratic relationship (*Quad*). In the similarity-based strategy class, we allowed exemplar models to include attention weights (ExAtt), to estimate additionally category or judgment biases (ExNoAttB and ExAttB), or to change the distance metric (ExDim). Subsuming a range of models in one strategy class, this strategy-class approach allows to determine the extent to which a participant can be best described by a rule-based strategy class or a similarity-based strategy class independent of the specific assumptions of one particular model. Further, we determined model fit and strategy classifications not only based on the generalization test, but also estimated the models' parameters based on both the training and the validation set (i.e. *fitting to training and test*). For this method, we penalized each model for model complexity by using Bayesian Information Criterion (BIC; Schwarz, 1978) and calculated BIC weights BICw_M for each strategy class in the strategy-class approach and the reduced set of strategies in the single-strategy approach. BICw_M can take values between 0 and 1. BICw_M close to 1

define a high posterior probability that the respective strategy class (or model) generated the data under the assumption that one of the considered classes is the data-generating model. In the following section, we determine whether varying the model selection technique and the set of strategy used may change the conclusions regarding the extent to which the task and the environment influence strategy classifications.

Results: Method Comparison

Predicting strategy choice. Figure 1 shows how average BIC and deviance weights for guessing, rule-based, and similarity-based strategies change depending on the model comparison technique (fitting to training and test or generalization) and strategies used (single strategy vs. strategy-class approach). As can be seen, most graphs show a similar pattern of results. BIC or deviance weights for guessing strategies are low in the OLIN, MLIN, and MMULT environment, but slightly increase in the MQUAD environment. Further, BIC and deviance weights in most graphs decrease for rule-based strategies from the OLIN to the MQUAD environment, whereas BIC and deviance weights for similarity-based strategies increase from the OLIN to the MQUAD environment. However, analyzing categorizations by fitting multiple strategies from the rule-based and similarity-based strategy class to training and test forms an exception: In the OLIN environment, this method shows higher evidence for similarity-based strategies than for rule-based strategies, whereas all other methods indicate more evidence for rule-based strategies.

Similarly, strategy classifications indicate that more participants are classified to guessing when relying on the single-strategy approach (see Table 6 and 7 for classifications). Further, most classification approaches indicate that more participants are classified to rule-based strategies in the OLIN and the MLIN environment than in the MMULT and MQUAD environment (with one exception in the OLIN categorization task). Lastly, slightly more participants are classified to similarity-based strategies in categorization than in judgment

(with the difference ranging from 8.6% using the generalization test in the single-strategy approach to 16.4% using fitting and a range of strategies).

To study the robustness of strategy classifications across different model comparison techniques and strategies used, we used a mixed logistic regression on strategy classifications from all four methods, but excluded participants best described by the guessing model separately for each task and method. Participants and method were included as random intercepts, whereas environment and task were included as fixed effects using the same contrasts for the environment as in the single-strategy analysis. Including random slopes did not change the results. Compared to a random model (AIC = 1199), a model that included the task domain, AIC = 1180, χ^2 (1) = 21.7, $p < .001$, provided a better fit. Including the environment as a main effect further improved model fit, AIC = 1151, χ^2 (3) = 34.2, $p < .001$, as did including the interaction between task domain and environment, AIC = 1113, χ^2 (3) = 43.9, $p < .001$. This final model showed a main effect of task domain suggesting that participants were less likely to rely on similarity-based strategies in judgment compared to categorization, OR = 0.31, CI = [0.19; 0.47] and a main effect of environment. Specifically, contrasts showed that more people were classified to rule-based strategies in linear environments than nonlinear environments, OR = 1.6, CI = [1.2; 2.2], but integrating more cues did not change strategy classification, OR = 0.92, CI = [0.61; 1.4]. In the nonlinear tasks, more participants were classified to similarity-based strategies in the MQUAD than in the MMULT environment, OR = 1.7, CI = [1.2; 2.6]. In addition, the model showed an interaction of task domain and environment. Breaking up the interaction by using contrasts indicated that differences in strategy choice between judgment and categorization were more pronounced in linear environments than in nonlinear environments, OR = 2.2; CI = [1.5; 3.8]. In the linear environments, more participants were classified to similarity-based strategies in categorization than in judgment in the OLIN task, but not in the MLIN task, OR = 6.2; CI = [3.1; 16.4]; in the nonlinear environments, the extent to which participants were classified to

rule- or similarity-based strategies in categorization and judgment did not vary between the MMULT or MQUAD environment, OR = 0.75; CI = [0.50; 1.1].

Strategy preferences and adaptation. Figure 4 shows the conditional probabilities for being assigned to a rule-based strategy class or strategy (or a similarity-based strategy class) in the second task given that participants were best described by a rule-based (or similarity-based) strategy class or strategy in the first task depending on model comparison technique (fitting to training and test or generalization) and strategies used (single-strategy approach vs. strategy-class approach). Overall, the different methods suggest a very similar pattern of findings with the main difference that fitting a variety of strategies to training and test suggests a lower probability of staying with rule-based strategies in the OLIN task than with the other methods used. Using the generalization test to discriminate between two strategy classes, on the other hand, suggests a higher probability of staying with similarity-based strategies in the MLIN task.

Discussion: Method Comparison

Applying different model comparison techniques or contrasting only a few candidate models may sometimes lead to different conclusions than considering a broader range of strategies or data (Donkin et al., 2015). Analyzing strategy choice across the different strategy classification methods, we found that accounting for method variance did not improve model fit — suggesting that overall strategy classifications did not vary strongly across different methods. Analyses with different model comparison techniques and sets of strategies replicated the result that in both judgment and categorization more participants are classified to rule-based strategies in linear than in nonlinear environments. Furthermore, considering all methods we found a small, but robust difference in classifications between the nonlinear tasks, with more participants classified to similarity-based strategies in the MQUAD environment than in the MMULT environment; the number of cues that need to be integrated, however, did not change strategy classifications across judgment and categorization. Furthermore,

compared to the single strategy, single method analysis, we found that more participants are better described by similarity-based strategies in categorization than in judgment — suggesting that the small advantage for similarity-based strategies in categorization is consistent across different methods. Finally, a closer analysis suggested an interaction of task domain and environment: more people were classified as following a similarity-based strategy in the OLIN environment in categorization than in the OLIN environment in judgment.

Taken together, a statistical analysis did not suggest that strategy classification varied across different model selection methods and sets of strategies, although eyeballing strategy weights and classifications indicated that some methods showed conflicting results, particularly in the OLIN environment. One reason why different methods may yield slightly different results is that methods can differ in regard to how reliably people can be classified to a specific strategy and how well a specific strategy can be recovered given noisy data. To ensure that the modeling approach we used allows meaningful classification we conducted a) a reliability analysis determining how reliable the classifications based on each approach are and b) a large model recovery study determining how often each model could be recovered by the generating strategy class using the different approaches (details are reported in Appendix B).

Overall, the reliability analysis suggested that strategy classifications show medium to high levels of reliability with single strategy classifications being slightly more reliable than strategy class classifications. The model recovery indicated that the majority of models could be recovered reasonably well in most environments and domains. One exception was the OLIN environment, in which exemplar models with attention weights could not be recovered in judgment. In categorization, the ability to recover a data-generating linear model or an exemplar model with attention weights depended on the model selection criterion and on the model generating the data. To ensure that problems to recover specific models did not influence our results in the OLIN environment, we designed a third experiment that allowed

us to better discriminate between rule-based and exemplar-based models in the OLIN environment (a detailed description is reported in Appendix C). Results from Experiment 3 replicated the results from Experiment 1, suggesting that a few more participants are classified to similarity-based strategies in categorization as compared to judgment, but employing different classification methods may lead to different estimates of the size of the effect. Taken together, these results reinforce the conclusion that differences between judgment and categorization are most pronounced in the OLIN environment.

General Discussion

The distinction between similarity-based and rule-based strategies is core to many areas of cognitive science (Hahn & Chater, 1998; Pothos, 2005; Sloman, 1996), but little research has linked similarity-based and rule-based strategies across different domains such as judgment and categorization. We have contributed to integrating judgment and categorization research by studying across a range of task characteristics and a range of instantiations of rule-based and similarity-based strategies whether the binary nature of feedback in categorization tasks more likely elicits similarity-based strategies than the continuous feedback people receive in judgment and how this depends on task characteristics, individual preferences, and the strategies considered.

Past literature jointly investigating categorization and judgment strategies has often argued that task feedback contributes to strategy changes between judgment and categorization tasks (Juslin, Olsson et al., 2003; Pachur & Olsson, 2012; von Helversen et al., 2010, 2013). We studied this hypothesis across a range of environments using two model comparison techniques and both a limited and a more exhaustive set of strategies. Across different methods, we found that less informative task feedback in categorization slightly increased shifts towards similarity-based strategies compared to continuous feedback in judgment. However, in relative terms only a small percentage of participants shifted from rule-based strategies in judgment to similarity-based strategies in categorizations. Further,

differences between domains were most pronounced in the one-dimensional linear task and less in a linear task with multiple cues or in nonlinear tasks. These results disagree with studies using binary cues reporting that more people rely on similarity-based strategies in categorization than judgment (Juslin, Olsson et al., 2003; Mata et al., 2012; von Helversen et al., 2010, 2013). One reason why the differences in feedback may have had less impact on processing in our study is that the quantitative cues we used implicitly conveyed knowledge about the cue directions (Newell et al., 2009). Knowledge about cue directions has been identified as a strong predictor for strategy choice and hence may have led to more people relying on rule-based strategies in the categorization task (Platzer & Bröder, 2013; von Helversen et al., 2013).

With respect to the question of how task characteristics affect strategy choice we found that in both judgment and categorization, the functional relationship between cues and criterion influenced strategy choice. Contrasting linear and nonlinear functions, we found that functions deviating from a linear form promoted similarity-based strategies and decreased the reliance on rule-based strategies. These results match past research suggesting that people frequently switch between strategies in both judgment (Hoffmann et al., 2014; Juslin et al., 2008; Karlsson et al., 2007) and categorization (Juslin, Jones et al., 2003; Rouder & Ratcliff, 2004) and resonate with studies showing the importance of the functional relationship in judgment (Hoffmann et al., 2013; Juslin et al., 2008; Karlsson et al., 2007). Further, our results add to the debate on how the form of the decision bound influences decision bound and similarity-based models (Olsson et al., 2006; Maddox & Ashby, 1993; McKinley & Nosofsky, 1995, 1996): Considering a variety of rule-based and similarity-based strategies and different model selection methods, we found that a few more participants were classified to similarity-based strategies in quadratic than in multiplicative tasks suggesting that increasingly nonlinear decision bounds are better described by similarity-based strategies, in both judgment and categorization.

Although past research has identified the number of cues as an important factor driving processing differences between one-dimensional tasks and tasks requiring information integration (Ashby et al., 2002; Filoteo et al., 2010; Maddox & Ashby, 2004; Zeithamova & Maddox, 2006), we found that the number of cues only impacted strategy choice in both judgment and categorization when comparing two representatives from each strategy class. When extending our analysis to a larger variety of strategies and model comparison methods, the number of cues, however, had a negligible impact on strategy choice across judgment and categorization. Instead, feedback altered the extent to which participants were classified to rule-based or similarity-based strategies in the linear environments. In the one-dimensional linear environment, more participants were classified as following similarity-based strategies in categorization than in judgment, but this difference was less pronounced in the linear task with multiple cues. One reason why more participants were classified to similarity-based strategies in the one-dimensional linear environment is possibly that we considered a broader range of strategies than in previous research. Variations in the strategies considered may alter to what extent the similarity-based strategy class can identify all exemplar users (Donkin et al., 2015). Accordingly, previous research may have underestimated the degree to which similarity may also influence categorization decisions in simple tasks.

As one of the first studies, we directly investigated the extent to which people adopt similar strategies in the domains of judgment and categorization. Specifically, we contrasted the idea that individual preferences underlie strategy choice with the idea that strategy use results from a slow adaptation to the environment. Overall, our results did not support the idea that people may have constant preferences for one strategy across tasks: Conditional probabilities that people relied on a similarity-based strategy in the second task given that they relied on a similarity-based strategy in the first task did not match the corresponding conditional probabilities for rule-based strategies. Further, the magnitudes systematically changed as a function of the environment. However, those participants who switched

strategies between tasks did not benefit from these switches compared to participants who transferred knowledge from the first to the second task. Instead, participants who switched solved both tasks less accurately, suggesting that they may not have succeeded in building up a coherent task representation.

In comparison to previous studies, we based strategy classifications upon different model comparison methods, fitting and generalization, considering both a range of strategies within a rule-based (or a similarity-based) strategy class as well as only two strategies per class. Although we obtained similar results applying those different approaches, the methods systematically differed in how well they recovered the models considered and in how consistently they classified participants to the same strategy class. Furthermore, the reliability analyses and model recovery study showed that the ability to recover a model depends also on the experimental task and its design. In particular, in simple tasks such as the OLIN environment data that can tease models apart may require a careful construction of the experiment in order not to bias the results of strategy classification to one of the models.

On a theoretical level, our study matches well with the idea that people can rely on both similarity-based and rule-based strategies (Erickson & Kruschke, 1998; Juslin, Olsson et al., 2003; Nosofsky et al., 1994; von Helversen & Rieskamp, 2008). Although our study conceptualized the interaction of rules and similarity as shifts between cognitive strategies, it is also likely that people base judgment and categorizations simultaneously on rules and similarity by blending these two cognitive strategies (Hahn, Prat-Sala, Pothos, & Brumby, 2010; von Helversen, Herzog, & Rieskamp, 2014). Future computational accounts may further exploit how rules and similarity interact by disentangling shifting and blending processes.

Taken together, our study suggests that people approach more complex tasks by relying more on similarity, with task complexity being a function of the functional relationship between the cues and the criterion and the number of the cues. However, whether

a task requires a categorization or a judgment and thus whether people receive binary or more fine-grained feedback influences strategy choice the most in simple linear tasks suggesting that the influence of feedback is overridden in more complex tasks. Strategy choice may be better understood as an adaptation process that is more challenging if the identification of task-appropriate strategies is more difficult. Studying how people deal with a range of cognitive problems may thus help to identify the conditions systematically triggering rule-based or similarity-based strategies.

Footnotes

1. Sometimes categorizations based on decision bounds are considered jointly with exemplar models as information integration models (Donkin et al., 2015), because the categorization likelihood depends on the distance of an item from the decision bound and is thus a question of similarity. Further, both exemplar and decision bound processes are often assumed to involve automatic processing (Ashby et al., 1998; Juslin et al., 2008). Here, we subsume decision bound models under rule-based processes for two reasons. First, categorizations based on decision bounds are made on the basis of abstracted knowledge even if that knowledge may not be consciously accessible, whereas exemplar-based categorizations rely on the comparison to exemplars from memory and thus a different cognitive process. Second, the level of explicit reasoning underlying the process is difficult to ascertain and in particular for linear rules evidence suggests that people have insight (Lagnado et al., 2006).

2. In order not to overweight tiny differences in model predictions, predicted percentages in categorization could not fall below .001. Similarly, the fitted standard deviations had to exceed .001.

References

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review, 105*, 442–481. doi:10.1037/0033-295X.105.3.442

Ashby, F. G., & Gott, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*, 35–53. doi:10.1037/0278-7393.14.1.33

Ashby, F. G., & Maddox, W. T. (1992). Complex decision rules in categorization: Contrasting novice and experienced performance. *Journal of Experimental Psychology: Human Perception and Performance, 18*, 50–71. doi:10.1037/0096-1523.18.1.50

Ashby, F. G., & Maddox, W. T. (2005). Human category learning. *Annual Review of Psychology*, *56*, 149–178. doi:10.1146/annurev.psych.56.091103.070217

Ashby, F. G., Maddox, W. T., & Bohil, C. J. (2002). Observational versus feedback training in rule-based and information-integration category learning. *Memory & Cognition, 30*, 666–677. doi:10.3758/BF03196423

Ashby, F. G., & O'Brien, J. B. (2005). Category learning and multiple memory systems. *Trends in Cognitive Sciences, 9,* 83–89. doi:10.1016/j.tics.2004.12.003

Ashby, F. G., Waldron, E. M., Lee, W. W., & Berkman, A. (2001). Suboptimality in human categorization and identification. *Journal of Experimental Psychology: General, 130,* 77-96. doi: 10.1037//0096-3445.130.1.77

Brehmer, B. (1994). The psychology of linear judgement models. *Acta Psychologica*, *87*, 137–154. doi:10.1016/0001-6918(94)90048-5

Brennan, R. L., & Prediger, D. J. (1981). Coefficient kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurement, 41*, 687-699.

Bröder, A., Glöckner, A., Betsch, T., Link, D., & Ettlin, F. (2013). Do people learn option or strategy routines in multi-attribute decisions? The answer depends on subtle factors. *Acta Psychologica, 143*, 200–209. doi:10.1016/j.actpsy.2013.03.005

Busemeyer, J. R., Byun, E., DeLosh, E. L., & McDaniel, M. A. (1997). Learning functional relations based on experience with input-output pairs by humans and artificial neural networks. In Lamberts K. & D. Shanks (Eds.), *Concepts and Categories* (pp. 405–437). Cambridge: MIT Press.

Busemeyer, J. R., & Wang, Y.-M. (2000). Model comparisons and model selections based on generalization criterion methodology. *Journal of Mathematical Psychology, 44*, 171–189. doi:10.1006/jmps.1999.1282

Cooksey, R. W. (1996). The Methodology of Social Judgement Theory. *Thinking & Reasoning, 2*, 141–174. doi:10.1080/135467896394483

DeLosh, E. L., Busemeyer, J. R., & McDaniel, M. A. (1997). Extrapolation: the sine qua non for abstraction in function learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 23, 968–986. doi:10.1037/0278-7393.23.4.968

Donkin, C., Newell, B. R., Kalish, M., Dunn, J. C., & Nosofsky, R. M. (2015). Identifying strategy use in category learning tasks: A case for more diagnostic data and models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*, 933–948. doi:10.1037/xlm0000083

Einhorn, H. J. (1971). Use of nonlinear, noncompensatory models as a function of task and amount of information. *Organizational Behavior and Human Performance, 6*, 1-27. doi:10.1016/0030-5073(71)90002-X

Elwin, E., Juslin, P., Olsson, H., & Enkvist, T. (2007). Constructivist coding: Learning from selective feedback. *Psychological Science, 18*, 105–110. doi:10.1111/j.1467-9280.2007.01856.x

Erickson, M. A., & Kruschke, J. K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General, 127*, 107–140. doi:10.1037/0096-3445.127.2.107

Filoteo, J. V., Lauritzen, J. S., & Maddox, W. T. (2010). Removing the frontal lobes: The effects of engaging executive functions on perceptual category learning. *Psychological Science, 21*, 415–423. doi:10.1177/0956797610362646

Hahn, U., & Chater, N. (1998). Similarity and rules: Distinct? Exhaustive? Empirically distinguishable? *Cognition, 65*, 197–230. doi:10.1016/S0010-0277(97)00044-9

Hahn, U., Prat-Sala, M., Pothos, E. M., & Brumby, D. P. (2010). Exemplar similarity and rule application. *Cognition, 114*, 1–18. doi:10.1016/j.cognition.2009.08.011

Hoffmann, J. A., von Helversen, B., & Rieskamp, J. (2013). Deliberation's blindsight: How cognitive load can improve judgments. *Psychological Science, 26*, 869-879. doi: 10.1177/0956797612463581

Hoffmann, J. A., von Helversen, B., & Rieskamp, J. (2014). Pillars of judgment: How memory abilities affect performance in rule-based and exemplar-based judgments. *Journal of Experimental Psychology: General, 143*, 2242–2261. doi:10.1037/a0037989

Juslin, P., Jones, S., Olsson, H., & Winman, A. (2003). Cue abstraction and exemplar memory in categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 924–941. doi:10.1037/0278-7393.29.5.924

Juslin, P., Karlsson, L., & Olsson, H. (2008). Information integration in multiple cue judgment: A division of labor hypothesis. *Cognition*, *106*, 259–298. doi:10.1016/j.cognition.2007.02.003

Juslin, P., Olsson, H., & Olsson, A.-C. (2003). Exemplar effects in categorization and multiple-cue judgment. *Journal of Experimental Psychology: General*, *132*, 133–156. doi:10.1037/0096-3445.132.1.133

Kämmer, J. E., Gaissmaier, W., & Czienskowski, U. (2013). The environment matters: Comparing individuals and dyads in their adaptive use of decision strategies. *Judgment and Decision Making, 8*, 299-329.

Karelaia, N., & Hogarth, R. M. (2008). Determinants of linear judgment: A meta-analysis of lens model studies. *Psychological Bulletin, 134*, 404–426. doi:10.1037/0033-2909.134.3.404

Karlsson, L., Juslin, P., & Olsson, H. (2007). Adaptive changes between cue abstraction and exemplar memory in a multiple-cue judgment task with continuous cues. *Psychonomic Bulletin & Review*, *14*, 1140–1146. doi:10.3758/BF03193103

Karlsson, L., Juslin, P., & Olsson, H. (2008). Exemplar-based inference in multi-attribute decision making: Contingent, not automatic, strategy shifts? *Judgment and Decision Making*, *3*, 244–260.

Kaufmann, E., & Athanasou, J. A. (2009). A meta-analysis of judgment achievement as defined by the lens model equation. *Swiss Journal of Psychology, 68*, 99–112. doi:10.1024/1421-0185.68.2.99

Lagnado, D. A., Newell, B. R., Kahan, S., & Shanks, D. R. (2006). Insight and strategy in multiple-cue learning. *Journal of Experimental Psychology: General, 135,* 162–183. doi:10.1037/0096-3445.135.2.162

Maddox, W. T., & Ashby, F. G. (1993). Comparing decision bound and exemplar models of categorization. *Perception & Psychophysics, 53*, 49–70. doi:10.3758/BF03211715

Maddox, W. T., & Ashby, F. G. (2004). Dissociating explicit and procedural-learning based systems of perceptual category learning. *Behavioural Processes, 66*, 309–332. doi:10.1016/j.beproc.2004.03.011

Maddox, W. T., Bohil, C. J., & Ing, A. D. (2004). Evidence for a procedural-learning-based system in perceptual category learning. *Psychonomic Bulletin & Review, 11*, 945–952. doi:10.3758/BF03196726

Maddox, W.T., Glass, B.D., Wolosin, S.M., Savarie, Z.R., Bowen, C., Matthews, M.D., & Schnyer, D.M. (2009). Sleep deprivation and information-integration: The effects of sleep deprivation on information-integration categorization performance. Sleep, 32, 1439-1448.

Mata, R., von Helversen, B., Karlsson, L., & Cüpper, L. (2012). Adult age differences in categorization and multiple-cue judgment. *Developmental Psychology*, *48*, 1188–1201. doi:10.1037/a0026084

McDaniel, M. A., & Busemeyer, J. R. (2005). The conceptual basis of function learning and extrapolation: Comparison of rule-based and associative-based models. *Psychonomic Bulletin & Review, 12,* 24–42. doi:10.3758/BF03196347

McDaniel, M. A., Cahill, M. J., Robbins, M., & Wiener, C. (2013). Individual differences in learning and transfer: Stable tendencies for learning exemplars versus abstracting rules. *Journal of Experimental Psychology: General.* Advance online publication. doi:10.1037/a0032963

McKinley, S. C., & Nosofsky, R. M. (1995). Investigations of exemplar and decision bound models in large, ill-defined category structures. *Journal of Experimental Psychology: Human Perception and Performance*, *21*, 128–148. doi: 10.1037/0096-1523.21.1.128

McKinley, S. C., & Nosofsky, R. M. (1996). Selective attention and the formation of linear decision boundaries. *Journal of Experimental Psychology: Human Perception and Performance*, *22*, 294–317. doi:10.1037/0096-1523.22.2.294

Markant, D. B., & Gureckis, T. M. (2014). Is it better to select or to receive? Learning via active and passive hypothesis testing. *Journal of Experimental Psychology: General, 143*, 94–122. doi:10.1037/a0032108

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review, 85*, 207–238. doi:10.1037/0033-295X.85.3.207

Medin, D. L., & Schwanenflugel, P. J. (1981). Linear Separability in Classification Learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 7*, 355–368. doi:10.1037/0278-7393.7.5.355

Newell, B. R., Weston, N. J., Tunney, R. J., & Shanks, D. R. (2009). The effectiveness of feedback in multiple-cue probability learning. *Quarterly Journal of Experimental Psychology*, *62*, 890–908. doi:10.1080/17470210802351411

Nosofsky, R. M. (1988). Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 700–708. doi:10.1037//0278-7393.14.4.700

Nosofsky, R. M., & Johansen, M. K. (2000). Exemplar-based accounts of "multiple-system" phenomena in perceptual categorization. *Psychonomic Bulletin & Review, 7*, 375–402.

Nosofsky, R. M., Little, D. R., & Denton, S. E. (2011). Response-time tests of logical-rule models of categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*, 1–27. doi:10.1037/a0021330

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review, 101*, 53–79. doi:10.1037/0033-295X.101.1.53

Nosofsky, R. M., & Zaki, S. R. (1998). Dissociations between categorization and recognition in amnesics and normal individuals: An exemplar-based interpretation. *Psychological Science, 9*, 247–255. doi:10.1111/1467-9280.00051

Olsson, A.-C., Enkvist, T., & Juslin, P. (2006). Go with the flow: How to master a nonlinear multiple-cue judgment task. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*, 1371–1384. doi: 10.1037/0278- 7393.32.6.1371

Olsson, H., Wennerholm, P., & Lyxzèn, U. (2004). Exemplars, prototypes, and the flexibility of classification models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*, 936–941. doi:10.1037/0278-7393.30.4.936

Pachur, T., & Olsson, H. (2012). Type of learning task impacts performance and strategy selection in decision making. *Cognitive Psychology*, *65*, 207–240. doi:10.1016/j.cogpsych.2012.03.003

Persson, M., & Rieskamp, J. (2009). Inferences from memory: Strategy- and exemplar-based inference models. *Acta Psychologica, 130,* 25-37. doi:10.1016/j.actpsy.2008.09.010

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review, 109*(3), 472–491. doi:10.1037//0033-295X.109.3.472

Platzer, C., & Bröder, A. (2013). When the rule is ruled out: Exemplars and rules in decisions from memory. *Journal of Behavioral Decision Making*, *26*, 429–441. doi: 10.1002/bdm.1776

Pothos, E. M. (2005). The rules versus similarity distinction. *The Behavioral and Brain Sciences*, *28*, 1–14. doi: 10.1017/S0140525X05000014

Rieskamp, J., & Otto, P. E. (2006). SSL: A theory of how people learn to select strategies. *Journal of Experimental Psychology: General,* 135, 207–236. doi:10.1037/0096-3445.135.2.207

Rouder, J. N., & Ratcliff, R. (2004). Comparing categorization models. *Journal of Experimental Psychology: General*, *133*, 63–82. doi:10.1037/0096-3445.133.1.63

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*, 461–464. doi:10.1214/aos/1176344136

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin, 119*, 3–22. doi:10.1037/0033-2909.119.1.3

von Helversen, B., Herzog, S. M., & Rieskamp, J. (2014). Haunted by a Doppelgänger: Irrelevant facial similarity affects rule-based judgments. *Experimental Psychology, 61*, 12-22. doi:10.1027/1618-3169/a000221

von Helversen, B., Karlsson, L., Mata, R., & Wilke, A. (2013). Why does cue polarity information provide benefits in inference problems? The role of strategy selection and knowledge of cue importance. *Acta Psychologica, 144*, 73–82. doi:10.1016/j.actpsy.2013.05.007

von Helversen, B., Mata, R., & Olsson, H. (2010). Do children profit from looking beyond looks? From similarity-based to cue abstraction processes in multiple-cue judgment. *Developmental Psychology, 46*, 220–229. doi:10.1037/a0016690

von Helversen, B., & Rieskamp, J. (2008). The mapping model: A cognitive theory of quantitative estimation. *Journal of Experimental Psychology: General,* 137, 73–96. doi:10.1037/0096-3445.137.1.73

von Helversen, B., & Rieskamp, J. (2009). Models of quantitative estimations: Rule-based and exemplar-based processes compared. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* 867–889. doi:10.1037/a0015501

Wagenmakers, E.-J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review, 11*, 192–196. doi:10.3758/BF03206482

Zeithamova, D., & Maddox, W. T. (2006). Dual-task interference in perceptual category learning. *Memory & Cognition, 34*, 387–398. doi:10.3758/BF03193416

Appendix A: Cognitive modeling of rule-based and similarity-based strategies

To identify the cognitive strategies that people rely on, we used a computational modeling approach. We compared how well the most prominent representative of the class of rule-based models, the linear model, described and predicted participants' responses in comparison to one often-used model from the class of similarity-based strategies, an exemplar model with one sensitivity parameter. In a second step, we contrasted this single-strategy account with a strategy class account that compares a class of rule-based strategies (using 5 different rule-based models) to the class of similarity-based strategies (using 5 different exemplar-based models).

Guessing models. To account for random guessing in categorization, we included a random guessing model assuming that participants' responses vary randomly around a mean on each trial (Maddox et al., 2009). For the judgment task, the guessing model estimated two free parameters: participants' mean judgment and the fitted standard deviation (see the section on model estimation); for the categorization task, it estimated participants' category bias (one free parameter).

Rule-based models. To model rule-based strategies, we fitted five different models of varying complexity. The linear model, corresponding mathematically to a linear regression model, has often served as the prototypical rule-based strategy in judgment tasks. It can represent simple rule-based strategies relying on a single cue, but also allows more complex rules combining several cues in a linear additive fashion. However, the linear model (Lin, 5 parameters) does not include nonlinear rules or interactions. Accordingly, the estimated criterion value $\hat{c}_p$ of an object p is the weighted sum of the cue values x_{pi},

$$\hat{c}_p = k + \sum_{i=1}^{4} w_i \cdot x_{pi} \tag{A1}$$

where w_i are the cue weights for each cue i and k is a constant intercept. Furthermore, we also fitted three models including nonlinear interactions between several cues. The optimal

multiplicative model (Mult, 7 parameters) to solve the multiplicative judgment task (Equation 3) included two three-way interaction terms:

$$\hat{c}_p = k + \sum_{i=1}^{4} w_i \cdot x_{pi} + w_5 \cdot \prod_{i=1}^{3} x_{pi} + w_6 \cdot \prod_{i=2}^{4} x_{pi} \quad \text{(A2)}$$

In addition, we also estimated two reduced multiplicative models (Mult1 and Mult2, 6 parameters) that only estimated one interaction term, for instance w_5, and sets the other one to zero, for instance w_6. Finally, we also fitted a quadratic model (Quad, 5 parameters) that centers each of the cue values at their mean. This model corresponds to a decision rule assuming that objects with high or low cue values have higher criterion values.

$$\hat{c}_p = k + \sum_{i=1}^{4} w_i \cdot \left(x_{pi} - 2.5\right)^2 \quad \text{(A3)}$$

For the categorization task we used these five rule-based models to determine the probability of classifying an object p with the estimated criterion value $\hat{c}$ to category b, by using a logistic classification function defined as: $p\left(\hat{b}=1\right) = \frac{e^{\hat{c}_p}}{1+e^{\hat{c}_p}}$. (A4)

The smoother logistic function accounts for random error in the decision-making process (Juslin, Jones et al., 2003).

Similarity-based models. To model similarity-based strategies, we fitted five different exemplar models. In exemplar models, the similarity $S(p,j)$ between the probe p and exemplar j is an exponential decay function of the distances d_{pj} between the objects (Nosofsky & Zaki, 1998).

$$S\left(p,j\right) = e^{-d_{pj}}. \quad \text{(A5)}$$

Thus, smaller distances between the probe p and exemplar j indicate a higher similarity between theses objects. To determine this distance, the cue values x_{pi} of probe p are compared to the cue values x_{ji} of exemplar j on all cues i. The more the cue values match each other, the smaller is the distance between the objects (Nosofsky & Johansen, 2000).

$$d_{pj} = h\left(\sum_{i=1}^{4} w_i \left|x_{pi} - x_{pj}\right|^r\right)^{1/r}. \tag{A6}$$

The sensitivity parameter *h* determines how strongly similarity decays with distance. Smaller sensitivity parameters indicate that similarity declines less with distance. The attention weights w_i, summing to one, weigh how much attention each cue or dimension receives. Finally, the distance metric is determined by *r*, with *r* = 1 typically used for separable dimensions and *r* = 2 for integral dimensions (Nosofsky & Johansen, 2000).

The probability of categorizing the probe *p* into response category *b*, $p\left(\hat{b}=1\right)$, can then be determined by calculating the similarity of probe *p* to all exemplars in category *b* and comparing it to the similarity of probe *p* to all exemplars (Nosofsky, 1988).

$$p\left(\hat{b}=1\right) = \frac{\sum_{j=1}^{J} \beta \cdot S\left(p, j_{b=1}\right)}{\sum_{j=1}^{J} \beta \cdot S\left(p, j_{b=1}\right) + \sum_{j=1}^{J} \left(1-\beta\right) \cdot S\left(p, j_{b=0}\right)} \tag{A7}$$

The category bias β captures how much people have a bias toward one of the two categories.

The simplest versions of the exemplar model only estimated the sensitivity parameter *h* (ExNoAtt, 1 parameter). The second model additionally assumed a category bias β (ExNoAttB, 2 parameters). Two more complex versions further estimated the attention weights, varying in whether they assumed a category bias (ExAttB, 5 parameters) or not (ExAtt, 4 parameters). Finally, the most complex model estimated the distance metric as well, but did not use a category bias (ExDim, 6 parameters). We did not include a response-scaling parameter because it can make exemplar models overly flexible (Olsson, Wennerholm, & Lyxzèn, 2004) and also does not find a corresponding mechanism in judgment. Accordingly, including an additional response-scaling parameter for exemplar models in categorization may bias model comparisons with exemplar models in judgment.

To account for judgments, Juslin, Olsson et al. (2003) assumed that the criterion value c_j of an exemplar is stored together with its cue values in memory. To estimate the criterion value of a new probe $\hat{c}_p$, the criterion values c_j for each exemplar are weighted by the similarities. Similar to the category bias in categorization, we also allowed participants' judgments to be biased towards high or low judgments by including an intercept *k*.

$$\hat{c}_p = k + \frac{\sum_{j=1}^{J} S(p,j) \cdot c_j}{\sum_{j=1}^{J} S(p,j)} \tag{A8}$$

We estimated five different versions of the exemplar model for the judgment tasks that closely mimicked the exemplar models for categorization. The simpler versions only estimated the sensitivity parameter (varying in whether they estimated the intercept, ExNoAttB, or not, ExNoAtt), the more complex versions further used attention weights (also using an intercept, ExAttB, or not, ExAtt). The most complex model also allowed the distance metric to vary without estimating an intercept (ExDim).

Model estimation and comparison. To evaluate the models' relative performance we used two model comparison techniques. In *fitting to training and test,* the models were fitted to both the last three training blocks and the four test blocks and evaluated based upon the Bayesian Information Criterion (BIC; Schwarz, 1978). Second, we compared model fits based upon a *generalization test* (Busemeyer & Wang, 2000) that uses the fitted parameters from the training phase to predict participants' responses in the test phase.

All judgment models were fitted to participants' responses by minimizing the deviance -2*LL*, the negative summed log-likelihood *L* of the model given the data.

$$-2LL = -2 \cdot \sum \ln(L) \tag{A9}$$

In the categorization task, the likelihood is defined as the models' predicted probability of the chosen category. In the judgment task, we calculated the likelihood as the probability density

of participants' judgments *j* assuming a truncated normal distribution, with the models' predicted responses $\hat{c}_p$ as the mean of the normal distribution and an estimated standard deviation σ.[1]

$$L = \frac{\frac{1}{\sigma}\phi\left(j\middle|\hat{c}_p,\sigma\right)}{\Phi\left(50\middle|\hat{c}_p,\sigma\right) - \Phi\left(0\middle|\hat{c}_p,\sigma\right)} \tag{A10}$$

To match the response scale from 0 to 50 we used this truncated normal distribution.

To compare which of the non-nested models described participants' responses better, we calculated the BIC for each model. This model selection criterion can be used to compare non-nested models. In addition, the BIC penalizes more complex models by accounting for the number of free model parameters *k*:

$$\text{BIC} = -2\text{LL} + k\ln(n), \tag{A11}$$

where *n* denotes the number of observations. Smaller BIC values indicate a better model fit. BICs were converted into BIC weights BICw_M that give the posterior probability of each model given the data relative to the competing models (Wagenmakers & Farrell, 2004):

$$\text{BIC}w_M = \frac{e^{-.5\Delta BIC_M}}{\sum_i e^{-.5\Delta BIC_i}} \tag{A12}$$

with ΔBIC_M as the difference between model *M* and the best model in the set and ΔBIC_i as the difference between a specific model *i* and the best model. For the *strategy-class approach*, BIC weights were summed for each set of rule-based and similarity-based strategies to yield the posterior probability of each strategy class (Donkin et al., 2015). For the *single-strategy approach*, we only included three models in the set (Guessing, Lin, and ExNoAtt) and derived BIC weights for each model based on the reduced set.

The parameter values, estimated for the training phase, were also used to predict participants' responses on the validation items of the test phase. To determine model fit, we determined the deviances based on the log likelihood of the predicted probability of the

observed responses. This generalization test corrects not only for model complexity in terms of the number of free parameters, but also for functional complexity (Busemeyer & Wang, 2000). We used the deviances further to calculate — in analogy to the BIC weights — deviance weights Dw_M for each model and each strategy class.

Appendix B: Reliability and model recovery for the different model selection criteria and strategy sets

Reliability analyses. The reliability analyses investigated how consistently participants are classified to rule-based or similarity-based strategies by each method when using different sets of data (cf. Brennan & Prediger 1981). To this goal we estimated odd-even reliabilities by splitting the test blocks into odd and even blocks. For fitting to training and test, the models were then fitted to the last three blocks of training and either the odd or the even test blocks. For the generalization test, the models were fit to the last three blocks of training and either predicted the odd or the even test blocks. We then calculated Cohen's κ using the percentage of concordant classifications based on BIC or deviance weights and corrected for the number of categories (strategy classes) because the number of observations per category was not equal (Brennan & Prediger, 1981).

Table B1 shows the odd-even reliabilities for the categorization task, separately for each environment, the model comparison technique (fitting to training and test vs. generalization test) and set of strategies used (strategy-class vs. single strategy); Table B2 depicts odd-even reliability for the judgment task. Descriptively, strategy classifications in the judgment task (κ = .77, *SD* = .16) are slightly more reliable than strategy classifications in the categorization task (κ = .74, *SD* = .15). Moreover, odd-even reliabilities decrease from the OLIN (κ = .92, *SD* = .08) to the MLIN (κ = .72, *SD* = .09) and MMULT (κ = .72, *SD* = .13) to the MQUAD environment (κ = .64, *SD* = .16). With regard to the classification method used, fitting to training and test (κ = .82, *SD* = .14) yields more reliable results than generalization (κ = .69, *SD* = .15) and the single-strategy approach (κ = .78, *SD* = .14) a higher odd-even reliability than the strategy-class approach (κ = .72, *SD* = .16). Fitting to training and test by using the single-strategy approach yields the highest reliability (κ = .87, *SD* = .11). In sum, an analysis of odd-even reliabilities suggests to only consider one rule-

based and one similarity-based strategy (as well as a guessing model) and to compare the models on their BIC values if model parameters are estimated using the training and test set.

Model recovery. The model recovery investigated whether we can identify the data-generating model using a specific set of strategies and different model comparison techniques (cf. Pitt, Myung, & Zhang, 2002).

To determine how often each model could be recovered by the generating strategy class using the different approaches, we drew 100 parameter sets for each model based on the best-fitting parameters for each individual from fitting the models to training and test, separately for each environment and the judgment and categorization task. We then determined response probabilities or average judgments for each of the 25 training and 15 validation items. Response probabilities were used to generate three simulated responses for each training item and four simulated responses for each validation item. Similarly, we used the average judgments to generate simulated judgments based on a truncated normal distribution with each model's median standard deviation.

To recover the models with fitting to training and test, we fitted all models (including a guessing model) to each simulated data set and calculated BIC weights for the single-strategy and the strategy-class approach. For the generalization test, we fitted all models to each simulated training set and calculated deviance weights for the single-strategy and the strategy-class approach based on the simulated data for the test set. A data set was deemed successfully recovered by the strategy-class approach if the generating strategy class had a higher summed BIC or deviance weight than the other strategy class or the guessing model. Similarly, for the single-strategy approach a model was deemed successfully recovered if its representative from the generating strategy class had a higher BIC or deviance weight than the guessing model or the representative from the other strategy class.

Figure B1 shows how often one model was successfully recovered by its strategy class in percent (black and white bars) as well as recovery by the single-strategy approach (gray

dots). The average percentage of models recovered per strategy class, environment, and approach used is presented in Table B1 for categorization and in Table B2 for judgment. Using the strategy-class approach, fitting to training and test (M = 85.0%, SE = 2.2%) on average recovered the models more successfully than the generalization test (M = 78.3%, SE = 2.3%). Using the single-strategy approach, fitting to training and test (M = 68.2%, SE = 3.7%) only slightly outnumbers the generalization test (M = 65%, SE = 3.2%). Further, the strategy-class approach also recovered a higher percentage of models than the single-strategy approach. Figure B1 also illustrates that the different approaches show a high variability in how successfully they recover different strategies in different environments. In the majority of the tasks model recovery was acceptable, with the exception of the OLIN task. Using fitting to training and test for OLIN categorizations the strategy-class approach successfully recovers exemplar models with attention weights, but does not recover rule-based strategies. Using the strategy-class approach with generalization, in contrast, does not recover exemplar models with attention weights for OLIN categorizations, but can reproduce rule-based strategies. Further, no approach successfully recovers exemplar models with attention weights for OLIN judgments.

In sum, we found that strategy classifications were most reliable when considering only a restricted set of strategies and fitting all strategies to training and test. However, this higher reliability comes at the cost of not detecting the specific data-generating model. The model recovery study showed that overall a strategy-class approach is more likely to recover the full set of models generating the data. In general, reliability and model recovery were reasonably high with exception of the OLIN environment, where in particular exemplar models with attention weights could not be recovered in the OLIN judgment task. This failure to recover some of the strategies might have biased the reported results for the OLIN environment. To rule out this conclusion, we designed a third experiment that allowed us to better discriminate between rule-based and exemplar-based models in the OLIN environment.

Table B1

Reliability (Percentage of Agreement, Cohen's κ*) and Model Recovery in the Categorization Task in the OLIN, MLIN, and MMULT Environment (Experiment 1) and in the MQUAD Environment (Experiment 2). Standard Error in Parentheses.*

	Strategy Class				Single Strategy			
	Reliability		Model Recovery		Reliability		Model Recovery	
Environment	%	κ	Rule	Sim	%	κ	Rule	Sim
				Fitting to Training and Test				
OLIN	100	1	50 (9)	94 (2)	97	.95	73 (18)	55 (18)
MLIN	78	.67	80 (10)	96 (1)	78	.67	74 (18)	84 (5)
MMULT	91	.86	87 (7)	98 (1)	94	.91	66 (17)	91 (4)
MQUAD	72	.58	46 (4)	86 (3)	91	.86	25 (7)	82 (4)
				Generalization Test				
OLIN	91	.86	82 (1)	47 (16)	91	.86	69 (17)	44 (17)
MLIN	75	.63	74 (3)	81 (2)	78	.67	68 (10)	69 (6)
MMULT	72	.58	79 (4)	80 (3)	72	.58	66 (12)	62 (8)
MQUAD	66	.48	63 (2)	80 (2)	75	.63	47 (9)	73 (6)

Note. OLIN = One-dimensional, linear environment; MLIN = Multidimensional, linear environment; MMULT = Multidimensional, multiplicative environment; MQUAD = Multidimensional, quadratic environment; Sim = Similarity

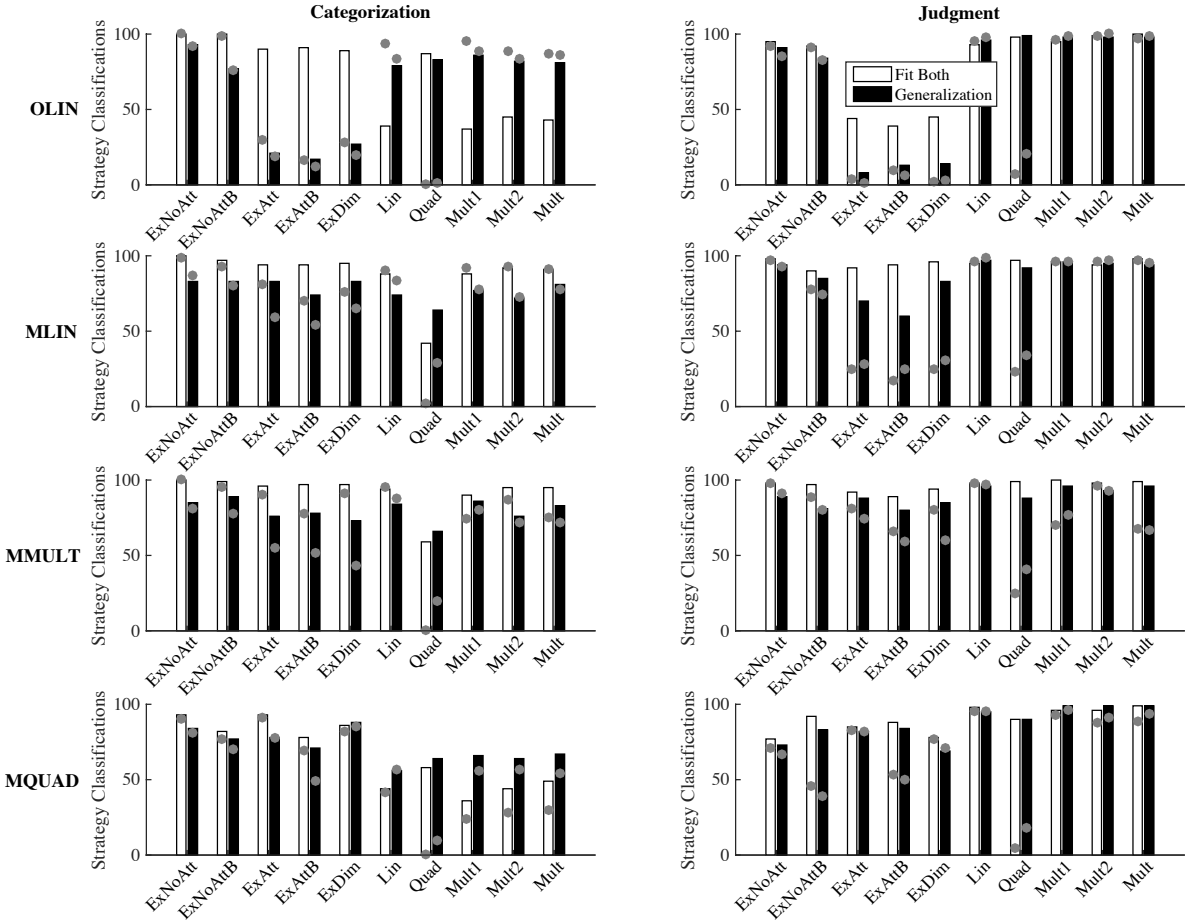
Table B2

Reliability (Percentage of Agreement, Cohen's κ*) and Model Recovery in the Judgment Task in the OLIN, MLIN, and MMULT Environment (Experiment 1) and in the MQUAD Environment (Experiment 2). Standard Error in Parentheses.*

	Strategy Class				Single Strategy			
	Reliability		Model Recovery		Reliability		Model Recovery	
Environment	%	κ	Rule	Sim	%	κ	Rule	Sim
				Fitting to Training and Test				
OLIN	88	.81	97 (1)	63 (13)	100	1	79 (18)	40 (21)
MLIN	84	.77	96 (1)	94 (1)	94	.91	82 (15)	48 (16)
MMULT	88	.81	99 (0)	94 (2)	84	.77	71 (13)	83 (5)
MQUAD	72	.58	96 (2)	84 (3)	94	.91	74 (17)	66 (7)
				Generalization Test				
OLIN	100	1	98 (0)	42 (19)	91	.86	83 (16)	36 (20)
MLIN	84	.77	95 (1)	78 (6)	81	.72	84 (13)	50 (14)
MMULT	72	.58	94 (2)	85 (2)	81	.72	75 (10)	73 (6)
MQUAD	69	.53	96 (2)	78 (3)	69	.53	79 (15)	62 (8)

Note. OLIN = One-dimensional, linear environment; MLIN = Multidimensional, linear environment; MMULT = Multidimensional, multiplicative environment; MQUAD = Multidimensional, quadratic environment; Sim = Similarity

Figure B1. Percentage of models recovered by their strategy class using two model comparison techniques, fitting to training and test (white bars) and the generalization test (black bars), separately for each task (columns), each environment (rows), and each generating model. Gray dots depict recovery by the single-strategy approach.



Appendix C: Discriminating between rule-based and exemplar-based strategies in one-dimensional environments

In Experiment 1, the model recovery suggested that exemplar models with attention weights are unlikely to be recovered in judgment and can only be recovered by fitting a broad range of strategies in categorization. To prevent more participants in categorization being classified to similarity-based strategies because of our inability to recover all similarity-based models in judgment, we conducted Experiment 3, which was designed to discriminate between rule-based and exemplar-based models by relying on fitting to training and test.

Method

Participants. Thirty-two participants (22 females, M_{Age} = 25.8, SD_{Age} = 4.6) were recruited from the same participant pool at the University of Basel as in Experiment 1. Participants from Experiment 1 and 2 were not allowed to take part in Experiment 3. Participants received course credit or a participation fee (20 CHF per hour). In addition, they could earn a bonus of 3 CHF in each task and had the opportunity to win one of two Amazon vouchers (worth 25 CHF each).

Design and material. We used the same cover stories and pictures as in Experiment 1. As in Experiment 1, the judgment criterion was a linear function of one cue, $y_{OLIN} = 10 * c_3$. To discriminate between the rule-based and the similarity-based strategy class, we selected the 25 training and 15 validation items so that the strategies made different predictions when fitting all strategies to training and test. We excluded items in the training set that required extrapolation to high cue values on c_3 because exemplar models should not be able to extrapolate beyond the range of training items in judgment. Further, we excluded items with cue values close to the boundary on c_3, because rule-based and similarity-based strategies should predict different responses on those items in categorization. Finally, to evaluate which method may discriminate best between the strategies, we conducted an a priori model recovery study using the same methodology as in Appendix B, but instead of using the fitted

parameters from each participant in the experiment, we used the median parameters from the OLIN environment in Experiment 1. Figure C1 shows the expected recovery rate for all strategies using different model selection criteria (fitting to training and test or generalization) for the strategy-class and the single-strategy approach. In the judgment task, fitting a range of rule-based and similarity-based strategies should allow good recovery of all strategies. For a few models, the expected percentage of recovered models is slightly lower in the categorization task, but fitting to training and test should still succeed in recovering most strategies.

Results

Accuracy in the judgment task. Judgment accuracy dropped from the last training block (M = 2.2, SD = 4.7) to accuracy in the test phase (M = 5.3, SD = 6.8). The order of the tasks did not influence judgment accuracy, $F(1,29) = 0.00$, $p = .981$, nor did the cover story, $F(1,29) = 0.03$, $p = .871$.

Accuracy in the categorization task. Similar to the judgment task, participants made on average few errors at the end of training (M = 8.0%, SD = 16.4), but the percentage of errors increased from training to test (M = 17.0%, SD = 18.3). Performance in the last training block was affected neither by order, $F(1,29) = 1.67$, $p = .206$, nor the cover story, $F(1,29) = 0.60$, $p = .444$.

Model fits in judgment. Table C1 shows BIC weights, deviance weights and strategy classifications in the judgment task. BIC weights and deviance weights were close to zero for the guessing model and no participant was best described by guessing. BIC weights from the strategy-class approach suggested that rule-based strategies outperformed the guessing model ($p < .001$, $r = -.92$), as did similarity-based strategies ($p < .001$, $r = -.76$). Further, rule-based strategies provided higher BIC weights than similarity-based strategies ($p < .001$, $r = -.65$). Results based on different methods led to similar conclusions.

Model fits in categorization. Like in the judgment task, BIC and deviance weights for the guessing model were small and the guessing model only described a few participants best (see Table C1). Both the rule-based strategies ($p < .001$, $r = -.87$) and similarity-based strategies ($p < .001$, $r = -.87$) outperformed the guessing model when comparing strategy classes on the basis of BIC weights. The BIC weights, however, did not differentiate rule-based from similarity-based strategies ($p = .667$, $r = -.08$). Using other model comparison methods, however, led to divergent results. The single-strategy approach suggested that rule-based strategies outperformed similarity-based strategies (fitting to training and test: $p < .001$, $r = -.83$; generalization: $p = .015$, $r = -.43$). In contrast, comparisons of the strategy-classes on the basis of a generalization test suggested that people are better described by similarity-based strategies ($p < .001$, $r = -.63$).

Predicting strategy choice. Descriptively, the majority of participants was classified to rule-based strategies in the judgment task, independent of the model comparison method or strategies used (see Table C1). In the categorization task, strategy classifications depended more strongly on the method used with a high number of participants classified to rule-based strategies using the single-strategy approach, but more participants classified to similarity-based strategies using the strategy-class approaches. This difference presumably stems from including exemplar models with attention weights in the strategy-class approach.

To analyze how the task, judgment vs. categorization, influenced strategy classifications, we again conducted a mixed logistic regression analysis on strategy classification in categorization and judgment, excluding participants classified to the guessing model. Participants and classification method were included as random intercept. Compared to a random model (AIC = 276), a model using task as a predictor performed better, AIC = 236, χ^2 (1) = 42.4, $p < .001$. Including a random slope for task further improved model fit, AIC = 233, χ^2 (2) = 6.5, $p = .038$. This model still suggested an overall impact of the task on strategy use with less people relying on similarity-based strategies in judgment than in

categorization, OR = 0.12, CI = [0.02; 0.79], p = .006, but the number of participants classified more to similarity-based strategies in categorization varied with the classification method.

Reliability. We calculated the odd-even reliabilities using the same approach as in Experiment 1 and 2. Table C2 lists the odd-even reliabilities separately for each task, model comparison technique, and set of strategies used. Across all methods, strategy classifications in judgment were more reliable (κ = .95, SD = .07) than in categorization (κ = .81, SD = .10). Further, generalization (κ = .92, SD = .10) yields slightly more reliable results than fitting to training and test (κ = .85, SD = .12) and the single-strategy approach a higher reliability (κ = .92, SD = .13) than using a broader range of strategies (κ = .85, SD = .08). In sum, the reliability analysis would suggest evaluating the strategy used based on a generalization test and a limited set of strategies.

Table C1

Model Fits and Classifications During Training and Test in the Categorization and Judgment Task in Experiment 3. Standard Error in Parentheses.

	Strategy class						Single strategy					
	Guessing		Rules		Similarity		Guessing		Rules		Similarity	
Task	M_w	n	M_w	n	M_w	n	M_w	n	M_w	n	M_w	n
						Fitting to training and test set						
Judgment	.00 (.00)	0	.78 (.07)	25	.22 (.07)	7	.00 (.00)	0	.97 (.03)	31	.03 (.03)	1
Categorization	.00 (.00)	0	.50 (.07)	18	.50 (.07)	14	.00 (.00)	0	.90 (.05)	29	.10 (.05)	3
						Generalization test						
Judgment	.00 (.00)	0	.94 (.04)	30	.06 (.04)	2	.00 (.00)	0	.94 (.04)	30	.06 (.04)	2
Categorization	.02 (.02)	1	.18 (.06)	6	.80 (.06)	25	.07 (.04)	2	.63 (.09)	20	.30 (.08)	10

Note. OLIN = One-dimensional, linear environment; M_w = Model weight

Table C2

Reliability (Percentage of Agreement, Cohen's κ*) in the Categorization and Judgment Task in the OLIN Environment (Experiment 3).*

Task	Strategy Class		Single Strategy	
	%	κ	%	κ
Fitting to Training and Test				
Judgment	91	.86	100	1
Categorization	88	.81	81	.72
Generalization Test				
Judgment	97	.95	100	1
Categorization	84	.77	97	.95

Note. OLIN = One-dimensional, linear environment; Sim = Similarity

Figure C1. Expected percentage of models recovered by their strategy class using fitting to training and test (white bars) or the generalization test (black bars), separately for each generating model. Gray dots depict expected recovery by the single-strategy approach.

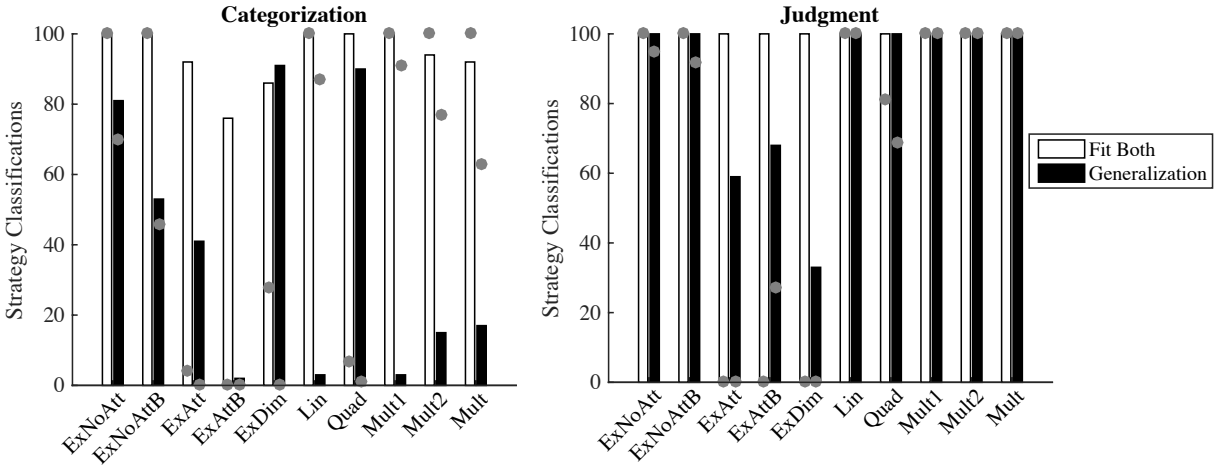


Table 1

Training Set for Study 1 Showing Cues, Judgment Criteria and Categorizations for the OLIN (derived from Equation 2), MLIN (Equation 1) and MMULT (Equation 3) Environment.

Cues				Judgment			Categorization		
Cue 1	Cue 2	Cue 3	Cue 4	OLIN	MLIN	MMULT	OLIN	MLIN	MMULT
2	1	0	3	0	14	2	0	0	0
1	4	1	4	10	22	5	0	0	0
0	3	1	2	10	13	2	0	0	0
0	2	3	0	30	12	1	1	0	0
5	5	4	0	40	43	29	1	1	1
0	4	5	4	50	26	12	1	1	1
2	4	3	0	30	26	9	1	1	1
1	4	3	5	30	27	13	1	1	1
1	0	2	4	20	12	1	0	0	0
1	0	0	2	0	6	1	0	0	0
5	3	3	5	30	40	21	1	1	1
1	1	5	5	50	22	7	1	0	1
1	2	0	5	0	15	2	0	0	0
5	5	0	1	0	36	4	0	1	0
0	4	3	1	30	19	4	1	0	0
4	2	1	3	10	27	6	0	1	1
0	5	2	3	20	22	6	0	0	1
5	5	2	4	20	43	22	0	1	1
5	1	3	4	30	33	9	1	1	1
4	0	2	4	20	24	3	0	0	0
1	4	1	5	10	23	6	0	0	1
3	0	5	5	50	27	3	1	1	0
0	2	5	0	50	16	2	1	0	0
1	5	2	4	20	27	10	0	1	1
3	4	5	5	50	39	30	1	1	1

Note. OLIN = One-dimensional, linear environment; MLIN = Multidimensional, linear environment; MMULT = Multidimensional, multiplicative environment

Table 2

Validation Set for Study 1 Showing Cues, Judgment Criteria and Categorizations for the OLIN (derived from Equation 2), MLIN (Equation 1) and MMULT (Equation 3) Environment.

Cues				Judgment			Categorization		
Cue 1	Cue 2	Cue 3	Cue 4	OLIN	MLIN	MMULT	OLIN	MLIN	MMULT
3	5	1	4	10	33	10	0	1	1
3	4	4	3	40	35	21	1	1	1
5	0	3	4	30	30	4	1	1	0
3	4	2	5	20	33	14	0	1	1
5	0	5	5	50	35	4	1	1	0
3	2	0	2	0	20	2	0	0	0
2	3	4	0	40	25	9	1	.5	1
4	5	4	5	40	44	36	1	1	1
5	0	5	3	50	33	4	1	1	0
4	3	0	1	0	26	3	0	1	0
2	1	2	0	20	15	3	0	0	0
2	5	2	3	20	30	12	0	1	1
4	0	0	2	0	18	2	0	0	0
4	1	1	1	10	22	4	0	0	0
3	3	3	5	30	32	15	1	1	1

Notes: OLIN = One-dimensional, linear environment; MLIN = Multidimensional, linear environment; MMULT = Multidimensional, multiplicative environment

Table 3

Training Set for Experiment 2. Judgment Criteria and Categorizations were Derived from Equation 4 for the Multidimensional Quadratic (MQUAD) Environment.

Cue 1	Cue 2	Cue 3	Cue 4	Judgment	Categorization
0	3	1	4	25	2
0	5	0	5	50	2
0	5	3	2	35	2
1	0	0	0	37	2
1	1	2	4	13	1
1	3	4	3	10	1
1	5	2	5	27	2
2	0	0	0	30	2
2	1	4	5	13	1
2	2	1	4	5	1
2	3	0	2	10	1
2	3	1	5	8	1
2	3	5	2	10	1
2	5	2	0	20	1
2	5	3	0	20	1
3	0	4	3	18	1
3	1	2	0	10	1
3	3	5	1	12	1
3	5	3	5	20	1
4	0	0	1	33	2
4	3	0	4	18	1
5	0	5	1	47	2
5	2	1	2	23	2
5	2	4	5	28	2
5	4	5	4	37	2

Table 4

Validation Set for Experiment 2. Judgment Criteria and Categorizations were Derived from Equation 4 for the Multidimensional Quadratic (MQUAD) Environment.

Cue 1	Cue 2	Cue 3	Cue 4	Judgment	Categorization
0	2	0	3	30	2
0	4	2	0	30	2
1	0	1	3	25	2
1	2	2	2	7	1
1	5	5	2	32	2
2	3	1	3	3	1
3	2	3	0	5	1
3	2	4	1	5	1
3	3	0	3	10	1
3	3	1	4	5	1
3	3	3	2	0	1
3	4	4	3	8	1
4	2	2	3	7	1
5	2	4	1	25	2
5	4	5	3	35	2

Table 5

Performance in the Judgment and Categorization Task in the OLIN, MLIN, and MMULT Environment (Experiment 1) and in the MQUAD Environment (Experiment 2). Standard Deviations in Parenthesis.

	Environment			
	OLIN	MLIN	MMULT	MQUAD
Categorization task				
% errors Training	3.8 (8.7)	22.5 (9.1)	23.4 (12.9)	29.3 (15.5)
% errors Test	3.5 (8.3)	24.0 (11.1)	21.8 (13.1)	35.2 (18.4)
Judgment task				
RMSD Training	4.2 (8.0)	6.7 (3.1)	5.4 (2.1)	11.9 (2.9)
RMSD Test	3.4 (6.2)	5.8 (1.5)	5.5 (1.9)	14.2 (2.6)

Note. OLIN = One-dimensional, linear environment; MLIN = Multidimensional, linear environment; MMULT = Multidimensional, multiplicative environment; MQUAD = Multidimensional, quadratic environment; RMSD = Root mean square deviation

Table 6

Model Weights and Strategy Classification in the Judgment Task in the OLIN, MLIN, and MMULT Environment (Experiment 1) and in the MQUAD Environment (Experiment 2). Standard Error for Model Weights in Parentheses.

	Strategy class						Single strategy					
	Guessing		Rules		Similarity		Guessing		Rules		Similarity	
Environment	M_w	n	M_w	n	M_w	n	M_w	n	M_w	n	M_w	n
Fitting to training and test set												
OLIN	.05 (.04)	2	.89 (.05)	29	.05 (.02)	1	.06 (.04)	2	.94 (.04)	30	.01 (.01)	0
MLIN	.03 (.02)	1	.73 (.08)	23	.24 (.07)	8	.04 (.03)	1	.77 (.07)	25	.18 (.07)	6
MMULT	.00 (.00)	0	.61 (.08)	19	.39 (.08)	13	.00 (.00)	0	.61 (.08)	20	.39 (.08)	12
MQUAD	.00 (.00)	0	.47 (.08)	15	.53 (.08)	17	.22 (.06)	7	.38 (.08)	12	.41 (.09)	13
Generalization test												
OLIN	.01 (.00)	0	.97 (.03)	31	.03 (.03)	1	.04 (.02)	1	.95 (.03)	31	.02 (.01)	0
MLIN	.01 (.01)	0	.55 (.08)	18	.44 (.08)	14	.03 (.02)	0	.64 (.08)	21	.33 (.08)	11
MMULT	.00 (.00)	0	.62 (.09)	20	.38 (.08)	12	.03 (.03)	1	.52 (.09)	16	.45 (.09)	15
MQUAD	.00 (.00)	0	.47 (.08)	16	.53 (.08)	16	.13 (.04)	3	.43 (.08)	15	.44 (.09)	14

Note. OLIN = One-dimensional, linear environment; MLIN = Multidimensional, linear environment; MMULT = Multidimensional, multiplicative environment; MQUAD = Multidimensional, quadratic environment; M_w = Model weight

Table 7

Model Weights and Strategy Classification in the Categorization Task in the OLIN, MLIN, and MMULT Environment (Experiment 1) and in the MQUAD Environment (Experiment 2). Standard Error in Parentheses.

	Strategy class						Single strategy					
	Guessing		Rules		Similarity		Guessing		Rules		Similarity	
Environment	M_w	n	M_w	n	M_w	n	M_w	n	M_w	n	M_w	n
Fitting to training and test set												
OLIN	.00 (.00)	0	.28 (.06)	8	.72 (.06)	24	.00 (.00)	0	.72 (.08)	23	.28 (.08)	9
MLIN	.02 (.02)	1	.69 (.07)	22	.29 (.07)	9	.02 (.02)	1	.81 (.06)	26	.17 (.06)	5
MMULT	.00 (.00)	0	.63 (.08)	20	.37 (.08)	12	.00 (.00)	0	.64 (.08)	21	.36 (.08)	11
MQUAD	.12 (.04)	3	.45 (.08)	14	.43 (.07)	15	.18 (.05)	5	.17 (.06)	6	.65 (.08)	21
Generalization test												
OLIN	.00 (.00)	0	.65 (.06)	25	.35 (.06)	7	.00 (.00)	0	.84 (.06)	27	.16 (.06)	5
MLIN	.04 (.03)	1	.56 (.08)	18	.40 (.08)	13	.05 (.03)	1	.54 (.08)	18	.41 (.08)	13
MMULT	.00 (.00)	0	.51 (.08)	15	.49 (.08)	17	.02 (.02)	1	.53 (.08)	17	.45 (.08)	14
MQUAD	.04 (.01)	0	.44 (.07)	13	.52 (.07)	19	.15 (.04)	4	.27 (.07)	9	.58 (.08)	19

Note. OLIN = One-dimensional, linear environment; MLIN = Multidimensional, linear environment; MMULT = Multidimensional, multiplicative environment; MQUAD = Multidimensional, quadratic environment; Mw = Model weight

Figure 1. BIC and deviance weights for the guessing, rule-based, and similarity-based strategies (lines in each plot), for judgment and categorization (rows). Big graphs depict results from the single-strategy approach using a generalization test. Smaller graphs show how results change across model comparison technique and strategies used. Error bars depict ± 1 standard error.

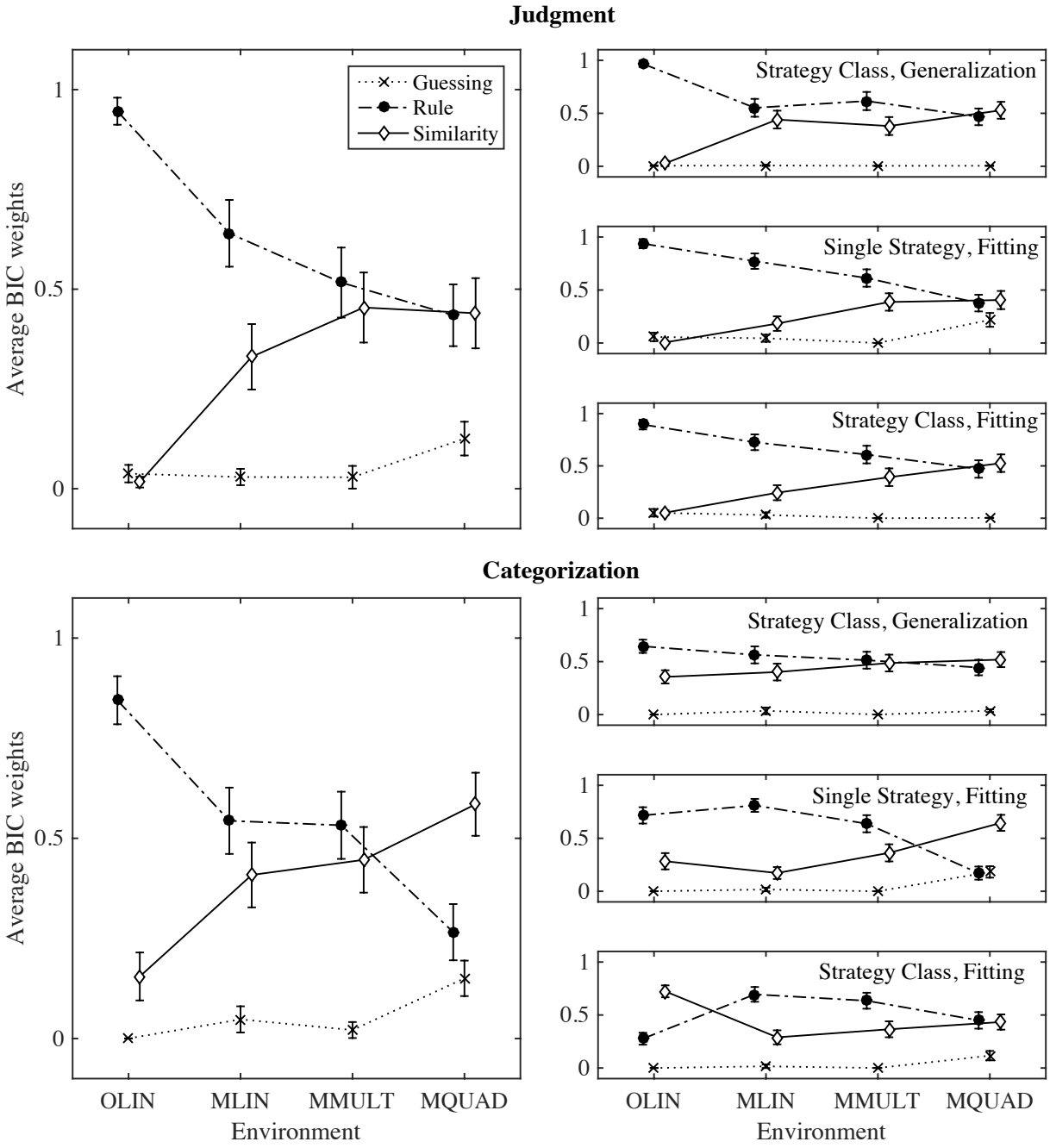


Figure 2. Model predictions from the linear model (white crosses, upper row) and ExNoAtt model (white crosses, lower row) averaged across those participants classified to the respective model, separately for the OLIN, MLIN, and MMULT (Experiment 1) and the MQUAD (Experiment 2) judgment task. Black diamonds represent average responses from participants classified to the linear model; gray circles represent average responses from participants classified to the ExNoAtt model. Black lines depict perfectly accurate judgments. In the OLIN environment, no participant was classified to the ExNoAtt model and model predictions hence based on a fit to the criterion.

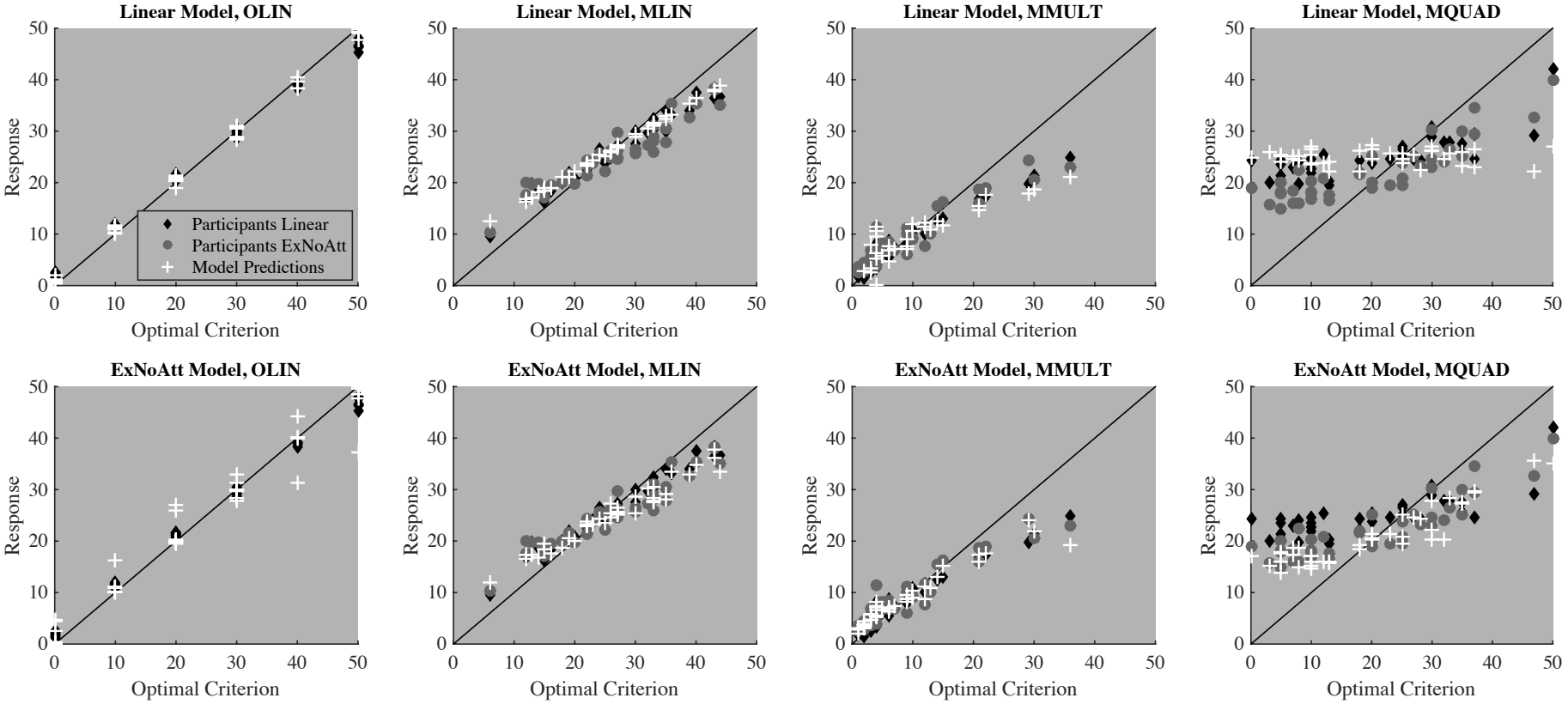


Figure 3. Model predictions from the linear model (white crosses, upper row) and the ExNoAtt model (white crosses, lower row) for those participants classified to the respective model, separately for the OLIN, MLIN, and MMULT (Experiment 1) and the MQUAD (Experiment 2) categorization task. Black diamonds represent responses probabilities from participants classified to the linear model, gray circles represent responses from participants classified to ExNoAtt model. Black lines depict median values that split the two categories.

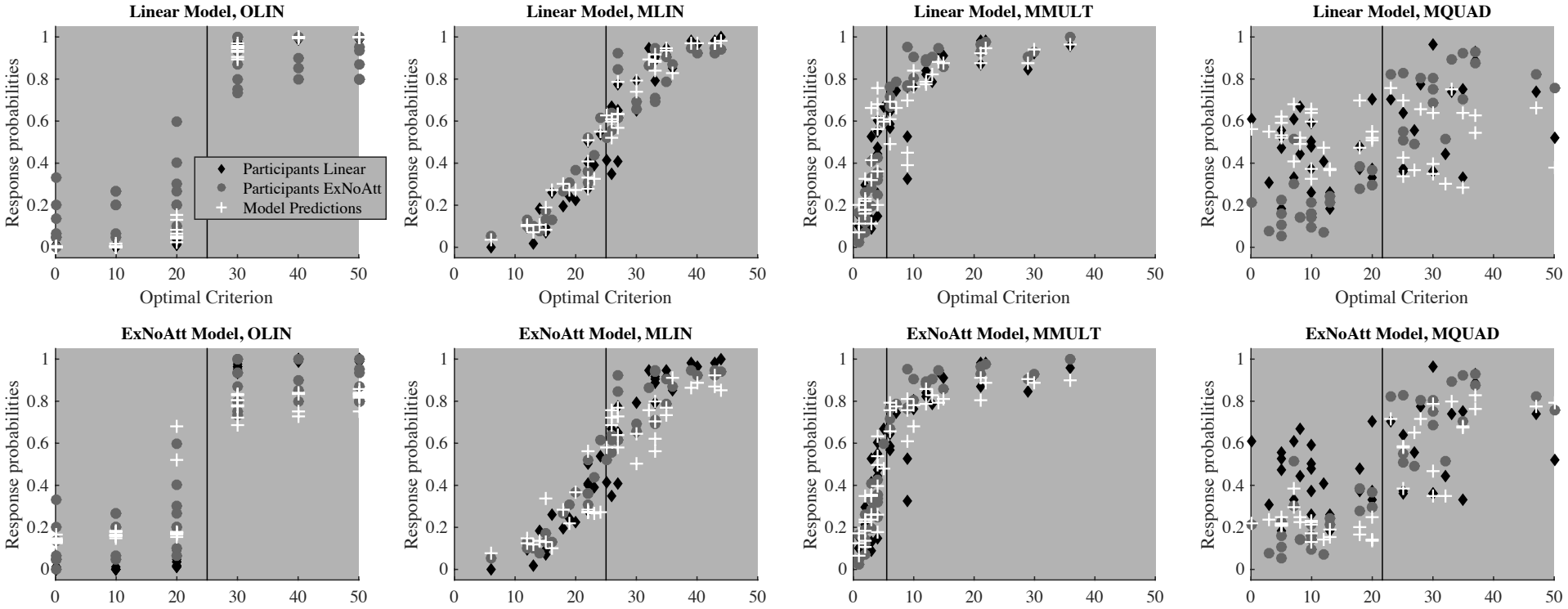


Figure 4. Conditional probabilities of classifying participants to the rule-based strategy class (white bars) or the similarity-based strategy class (gray bars) in the second task given that the participant was classified to rule-based or similarity-based strategies in the first task, respectively. Conditional probabilities are depicted for the OLIN (one-dimensional, linear), the MLIN (multidimensional, linear), and the MMULT (multidimensional, multiplicative) environment from Experiment 1 as well as for the MQUAD (multidimensional, quadratic) environment from Experiment 2. Big graphs depict results from the single-strategy approach using a generalization test. Smaller graphs show how results change across model comparison technique and strategies used.